PAC-BAYESIAN INDUCTIVE AND TRANSDUCTIVE LEARNING

Olivier Catoni

May 28, 2006

Abstract: We present here a PAC-Bayesian point of view on adaptive supervised classification. Using convex analysis on the set of posterior probability measures on the parameter space, we show how to get local measures of the complexity of the classification model involving the relative entropy of posterior distributions with respect to Gibbs posterior measures. We then discuss relative bounds, comparing the generalization error of two classification rules, showing how the margin assumption of Mammen and Tsybakov can be replaced with some empirical measure of the covariance structure of the classification model. We also show how to associate to any posterior distribution an *effective temperature* relating it to the Gibbs prior distribution with the same level of expected error rate, and how to estimate this effective temperature from data, resulting in an estimator whose expected error rate converges according to the best possible power of the sample size adaptively under any margin and parametric complexity assumptions. Then we introduce a PAC-Bayesian point of view on transductive learning and use it to improve on known Vapnik's generalization bounds, extending them to the case when the sample is made of independent non identically distributed pairs of patterns and labels. Eventually we review briefly the construction of Support Vector Machines and show how to derive generalization bounds for them, measuring the complexity either through the number of support vectors or through transductive or inductive margin estimates.

2000 Mathematics Subject Classification: 62H30, 68T05, 62B10.

Keywords: Statistical learning theory, adaptive statistics, pattern recognition, PAC-Bayesian theorems, VC dimension, local complexity bounds, randomized estimators, Gibbs posterior distributions, effective temperature, Mammen and Tsybakov margin assumption, transductive inference, compression schemes, Support Vector Machines, margin bounds.

CNRS – Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6 (site Chevaleret), 4 place Jussieu – Case 188, 75 252 Paris Cedex 05.

Contents

Introduction

Among the possible approaches to pattern recognition, statistical learning theory has received a lot of attention in the last few years. Although a realistic pattern recognition scheme involves data pre-processing and post-processing that need a theory of their own, a central role is often played by some kind of supervised learning algorithm. This central piece of work is the subject we are going to analyse in these notes.

Accordingly, we assume that we have prepared in some way or another a *sample* of N labelled patterns $(X_i, Y_i)_{i=1}^N$, where X_i ranges in some pattern space $\mathcal{X}$ and Y_i ranges in some finite label set $\mathcal{Y}$. We also assume that we have devised our experiment in such a way that the couples of random variables (X_i, Y_i) are independent (but not necessarily equidistributed). Here, randomness should be understood to come from the way the statistician has planned his experiment. He may for instance have drawn the X_is at random from some larger population of patterns the algorithm is meant to be applied to in a second stage. The labels Y_i may have been set with the help of some external expertise (which may itself be faulty or contain some amount of randomness, therefore we do not assume that Y_i is a function of X_i, and allow the couple of random variables (X_i, Y_i) to follow any kind of joint distribution). In practice, patterns will be extracted from some high dimensional and highly structured data, like digital images, speech signals, DNA sequences, etc. We will not discuss here this pre-processing stage (although it poses crucial problems dealing with segmentation and the choice of a representation).

To fix notations, let $(X_i, Y_i)_{i=1}^N$ be the canonical process on $\Omega = (\mathcal{X} \times \mathcal{Y})^N$ (which means the coordinate process). Let the pattern space be provided with a sigma-algebra $\mathcal{B}$ turning it into a measurable space $(\mathcal{X}, \mathcal{B})$. On the finite label space $\mathcal{Y}$, we will consider the trivial algebra $\mathcal{B}'$ made of all its subsets. Let $\mathcal{M}_+^1\bigl[(\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\bigr]$ be our notation for the set of probability measures (i.e. of positive measures of total mass equal to 1) on the measurable space $\bigl[(\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B} \times \mathcal{B}')^{\otimes N}\bigr]$. Once some probability distribution $\mathbb{P} \in \mathcal{M}_+^1\bigl[(\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\bigr]$ is chosen, it turns $(X_i, Y_i)_{i=1}^N$ into the canonical realization of a stochastic process modeling the observed sample (also called the training set). We will assume that $\mathbb{P} = \bigotimes_{i=1}^N P_i$, where for each $i = 1, \dots, N$, $P_i \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{B}')$, to reflect the assumption that we observe independent pairs of patterns and labels. We will also assume that we are provided with some indexed set of possible classification rules

$$\mathcal{R}_\Theta = \bigl\{ f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta \bigr\},$$

where $(\Theta, \mathcal{T})$ is some measurable index set. Assuming some indexation of the classification rules is just a matter of presentation. Although it leads to longer notations, it allows to integrate over the space of classification rules as well as over Ω using the usual formalism of multiple integrals. For this matter, we will assume that $(\theta, x) \mapsto f_\theta(x) : (\Theta \times \mathcal{X}, \mathcal{B} \otimes \mathcal{T}) \rightarrow (\mathcal{Y}, \mathcal{B}')$ is a measurable function.

In many cases $\Theta = \bigcup_{i \in I} \Theta_i$ will be a finite (or more generally countable) union of subspaces, dividing the classification model $\mathcal{R}_\Theta = \bigcup_{i \in I} \mathcal{R}_{\Theta_i}$ into a union of submodels. The importance of introducing such a structure has been put forward by V. Vapnik, as a way to avoid making strong hypotheses on the distribution $\mathbb{P}$ of the sample. If neither the distribution of the sample nor the set of classification rules were constrained, it is well known indeed that no kind of statistical inference would be possible. Considering a family of submodels is a way to provide for adaptive classification where the choice of the model depends on the observed sample. Restricting the set of classification rules is more realistic than restricting the distribution of patterns, since the classification rules are a processing tool left to the choice of the statistician, whereas the distribution of the patterns is not fully under his control, except for some planning of the learning experiment which may enforce some weak properties like independence, but not the precise shapes of the marginal distributions P_i which are as a rule unknown distributions on some high dimensional space.

In these notes, we will concentrate on general issues concerned with a natural measure of risk, namely the *expected error rate* of each classification rule f_θ, expressed as

$$R(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\bigl[f_\theta(X_i) \neq Y_i\bigr].$$

As this quantity is unobserved, we will be led to work with the corresponding *empirical error rate*

$$r(\theta, \omega) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\bigl[f_\theta(X_i) \neq Y_i\bigr].$$

This does not mean that pratical learning algorithms will always try to minimize this criterion. They often on the contrary try to minimize some other criterion which is linked with the structure of the problem and has some nice additional properties (like smoothness and convexity, for example). Nevertheless, and independently from the precise form of the estimator

$\widehat{\theta} : \Omega \to \Theta$ under study, the analysis of $R(\widehat{\theta})$ is a natural question, and often corresponds to what is required in practice.

Answering this question is not straightforward because, although $R(\theta)$ is the expectation of $r(\theta)$, a sum of independent Bernoulli random variables, $R(\widehat{\theta})$ is not the expectation of $r(\widehat{\theta})$, because of the dependence of $\widehat{\theta}$ on the sample, and neither is $r(\widehat{\theta})$ a sum of independent random variables. To circumvent this unfortunate situation, some uniform control over the deviations of r with respect to R is needed.

The PAC-Bayesian approach to this problem, originated in the machine learning community and pioneered by D. McAllester [25, 26], can be seen as some variant of the more classical approach of M-estimators relying on empirical process theory (as exposed for instance in [36]).

It is built on three corner stones:

- One idea is to embed the set of estimators of the type $\widehat{\theta} : \Omega \to \Theta$ into the larger set of regular conditional probability measures $\rho : \bigl(\Omega, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\bigr) \to \mathcal{M}_+^1(\Theta, \mathcal{T})$. We will call these conditional probability measures *posterior distributions*, to follow a usual terminology.
- A second idea is to measure the fluctuations of ρ with respect to the sample, using some prior distribution $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$, and the Kullback divergence function $\mathcal{K}(\rho, \pi)$. The expectation $\mathbb{P}\bigl\{\mathcal{K}(\rho, \pi)\bigr\}$ measures the randomness of ρ. The optimal choice of π would be $\mathbb{P}(\rho)$, resulting in a measure of the randomness of ρ equal to the mutual information between the sample and the estimated parameter drawn from ρ. Anyhow, since $\mathbb{P}(\rho)$ is as a rule no more observed than $\mathbb{P}$, we will have to be content with some less concentrated prior distribution π, resulting in some looser measure of randomness, as shown by the identity $\mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr] = \mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \mathbb{P}(\rho)\bigr]\bigr\} + \mathcal{K}\bigl[\mathbb{P}(\rho), \pi\bigr]$.
- A third idea is to analyze the fluctuations of the random process $\theta \mapsto r(\theta)$ with respect to its mean process $\theta \mapsto R(\theta)$ through the log-Laplace transform

$$-\frac{1}{\lambda} \log \left\{ \iint \exp\bigl[-\lambda r(\theta, \omega)\bigr] \pi(d\theta) \mathbb{P}(d\omega) \right\},$$

  as a physicist prone to statistical mechanics (where this is called the free energy) would do. This transform is well suited to relate $\min_{\theta \in \Theta} r(\theta)$ to $\inf_{\theta \in \Theta} R(\theta)$.

This monograph is devided into two sections. The first one deals with the inductive setting presented in these lines, the second one with the *transductive* setting, where, following Vapnik's seminal approach [37], a shadow

sample is considered.

In the first section, two types of bounds are shown. *Empirical bounds* can be used to choose between estimators or to build estimators. *Non random bounds* can be used to assess the speed of convergence of estimators, relating this speed to the speed of convergence of the Gibbs prior expected error rate $\beta \mapsto \pi_{\exp(-\beta R)}(R)$ towards $\operatorname{ess\,inf}_\pi R$ as β goes to infinity, and to other quantities akin to the margin assumption of Mammen and Tsybakov in more sophisticated cases. We will progress from the most straighforward bounds to more elaborate ones, built to achieve a better asymptotic behaviour. We will thus introduce *local bounds* and *relative bounds*. From an asymptotic point of view, the culminating result of these notes is Theorem 1.39 (page 63). It is used in Proposition 1.40 to build a classification rule which is proved to be adaptive in all the parameters of the Mammen and Tsybakov margin assumption and of a parametric complexity assumption in Corollary 1.52 (page 78) of Theorem 1.50 (page 77). This opens the road to Theorem 1.59 (page 88) which performs two step localization on top of Theorem 1.39 in order to be able to achieve adaptive model selection with a decreased influence of the number of empirically unefficient models included in the comparison. The analysis of this bound is hinted at in subsequent pages, but not fully developed, since we are not sure the amount of technicalities it requires is worth it. Anyhow we would not like to induce the reader into thinking that each result in the first section is actually an *improvement* on the previous one, it is as a rule only an *asymptotic improvement*, and the price to pay for being asymptotically tighter is to get looser bounds for small sample sizes. What is a small sample size in practice is a question of ratio between the number of examples and the complexity (roughly speaking the number of parameters) of the model used to classify. Since our aim here is to describe classification methods suitable for complex data (images, speech, DNA, ...), we suspect that practitioners wanting to make use of our proposals will be confronted with small sample sizes more often than with large ones, and should try to make use of the simplest bounds first and see only afterwards whether the asymptotically better ones can bring them more for the size of samples their computers can handle and their data bases can provide. Let us advocate also that the results of this first section are not only of a theoretical nature for two reasons : the first one is that posterior parameter distributions can be computed effectively, using Monte Carlo techniques, there is a whole tradition about these computations in Bayesian statistics, proving that what we call here Gibbs estimators are not only a way to show that some optimal speeds of convergence can be reached in some theoretically well understood situations, but that they can

also be computed in practice. The second reason is that a traditional non randomized estimator $\widehat{\theta} \in \Theta$ of the parameter can be approximated by a posterior distribution ρ which is supported by a fairly narrow neighboorhood of $\widehat{\theta} \in \Theta$, without spoiling excessively our bounds, resulting in a classification rule which is to provide a randomized answer only for a small amount of dubious examples and will most of the time issue the same deterministic answer as the classification rule indexed by $\widehat{\theta}$ it is derived from. This is explained on page 14.

In the second section, we show first how we can transport all the results obtained in the inductive case to the transductive case, allowing to replace prior distributions by *partially exchangeable posterior distributions* depending on an extended sample were unlabelled shadow examples are added, with increased possibilities of adaptation to the data. We then focus on the small sample case, where local and relative bounds are not expected to be of great help. Using a fictitious (that is unobserved) shadow sample, we study Vapnik type generalization bounds, showing how to tighten and extend them using some original ideas, like making no Gaussian approximation to the log-Laplace of Bernoulli random variables, — using a shadow sample of arbitrary size, — shrinking from the use of any symmetrization trick — and using a subset of the group of permutations suitable to cover the case of independent non identically distributed data. The culminating result of the second section is Theorem 2.17 on page 114, subsequent bounds showing the separate influence of the above ideas and providing an easier comparison with Vapnik's original results. Vapnik type generalization bounds have a broad applicability, not only through the concept of VC dimension, but also through the use of compression schemes [24], which are briefly described on page 105.

1. Inductive PAC-Bayesian learning

The setting of inductive inference (as opposed to transductive inference to be discussed later) is the one described in the introduction.

When we will have to take the expectation of a random variable $Z : \Omega \rightarrow \mathbb{R}$ as well as of a function of the parameter $h : \Theta \rightarrow \mathbb{R}$ with respect to some probability measure, we will as a rule use functional short notations instead of resorting to the integral sign: thus we will write $\mathbb{P}(Z)$ for $\int_\Omega Z(\omega)\mathbb{P}(d\omega)$ and $\pi(h)$ for $\int_\Theta h(\theta)\pi(d\theta)$.

The PAC-Bayesian approach, in its simplest form, relies on some basic upper bound for the Laplace transform of $\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \bigl[\rho(R) - \rho(r)\bigr]$, or more

technically on some penalized variant of it, as will be seen. This will be the subject of the next subsection, where we will start with the Laplace transform of $R(\theta) - r(\theta)$, for any $\theta \in \Theta$, before encompassing posterior distributions. As it is already easy to guess, the purpose of these preliminaries is to gain some uniform control on the lower deviations of the empirical error rate from the expected error rate under any posterior distribution.

1.1. Basic inequality. In the setting described in the introduction, let us consider the Bernoulli random variables $\sigma_i(\theta) = \mathbb{1}\bigl[Y_i \neq f_\theta(X_i)\bigr]$. Using independence and the concavity of the logarithm function, it is readily seen that for any real constant λ

$$\log\Bigl\{\mathbb{P}\bigl\{\exp\bigl[-\lambda r(\theta)\bigr]\bigr\}\Bigr\} = \sum_{i=1}^N \log\Bigl\{\mathbb{P}\Bigl[\exp\bigl(-\tfrac{\lambda}{N}\sigma_i\bigr)\Bigr]\Bigr\}$$
$$\leq N \log\Biggl\{\frac{1}{N}\sum_{i=1}^N \mathbb{P}\Bigl[\exp\bigl(-\tfrac{\lambda}{N}\sigma_i\bigr)\Bigr]\Biggr\}.$$

The right-hand side of this inequality is the log Laplace transform of a Bernoulli distribution with parameter $\frac{1}{N}\sum_{i=1}^N \mathbb{P}(\sigma_i) = R(\theta)$. As any Bernoulli distribution is fully defined by its parameter, this log Laplace transform is necessarily a function of $R(\theta)$. It can be expressed with the help of the family of functions

$$\Phi_a(p) = -a^{-1}\log\bigl\{1 - \bigl[1 - \exp(-a)\bigr]p\bigr\}, \quad a \in \mathbb{R}, p \in (0,1).$$

It is immediately seen that Φ_α is an increasing one to one mapping of the unit interval unto itself, and that it is convex when $a > 0$, concave when $a < 0$ and can be defined by continuity to be the identity when $a = 0$. Moreover the inverse of Φ_a is given by the formula

$$\Phi_a^{-1}(q) = \frac{1 - \exp(-aq)}{1 - \exp(-a)}, \qquad a \in \mathbb{R}, q \in (0,1).$$

This formula may be used to extend Φ_a^{-1} to $q \in \mathbb{R}$, and we will use this extension without further notice when required.

Using these notations, the previous inequality becomes

$$\log\Bigl\{\mathbb{P}\bigl\{\exp\bigl[-\lambda r(\theta)\bigr]\bigr\}\Bigr\} \leq -\lambda\Phi_{\frac{\lambda}{N}}\bigl[R(\theta)\bigr], \quad \text{proving}$$

LEMMA 1.1. *For any real constant λ and any parameter $\theta \in \Theta$,*

$$\mathbb{P}\Bigl\{\exp\Bigl\{\lambda\Bigl[\Phi_{\frac{\lambda}{N}}\bigl[R(\theta)\bigr] - r(\theta)\Bigr]\Bigr\}\Bigr\} \leq 1.$$

In previous versions of this study, we had used some Bernstein bound, instead of this lemma. Anyhow, as it will turn out, keeping the log Laplace of a Bernoulli instead of approximating it provides simpler and tighter results.

Lemma 1.1 implies that for any constants $\lambda \in \mathbb{R}_+$ and $\epsilon \in)0, 1)$,

$$\mathbb{P}\biggl[\Phi_{\frac{\lambda}{N}}\bigl[R(\theta)\bigr] + \frac{\log(\epsilon)}{\lambda} \leq r(\theta)\biggr] \geq 1 - \epsilon.$$

Choosing $\overline{\lambda} \in \arg\max_{\mathbb{R}_+} \Phi_{\frac{\lambda}{N}}\bigl[R(\theta)\bigr] + \dfrac{\log(\epsilon)}{\lambda}$, we deduce

LEMMA 1.2. *For any $\epsilon \in)0, 1)$, any $\theta \in \Theta$,*

$$\mathbb{P}\biggl\{R(\theta) \leq \inf_{\lambda \in \mathbb{R}_+} \Phi^{-1}_{\frac{\lambda}{N}}\biggl[r(\theta) - \frac{\log(\epsilon)}{\lambda}\biggr]\biggr\} \geq 1 - \epsilon.$$

We will illustrate throughout these notes the bounds we prove with a small numerical example: in the case where $N = 1000$, $\epsilon = 0.01$ and $r(\theta) = 0.2$, we get with a confidence level of 0.99 that $R(\theta) \leq .2402$, this being obtained for $\lambda = 234$.

Now, to proceed towards the analysis of posterior distributions, let us put for short $U_\lambda(\theta, \omega) = \lambda\Bigl[\Phi_{\frac{\lambda}{N}}\bigl[R(\theta)\bigr] - r(\theta, \omega)\Bigr]$, and let us consider $\log\Bigl\{\mathbb{P}\Bigl[\pi\bigl[\exp(U_\lambda)\bigr]\Bigr]\Bigr\}$, where $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$ is some prior probability measure on the parameter space. Using Fubini's theorem for non negative functions, we see that

$$\log\Bigl\{\mathbb{P}\Bigl[\pi\bigl[\exp(U_\lambda)\bigr]\Bigr]\Bigr\} = \log\Bigl\{\pi\Bigl[\mathbb{P}\bigl[\exp(U_\lambda)\bigr]\Bigr]\Bigr\} \leq 0.$$

To relate this quantity to the expectation $\rho(U_\lambda)$ with respect to any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, we will use the properties of the Kullback divergence $\mathcal{K}(\rho, \pi)$ of ρ with respect to π, which is defined as

$$\mathcal{K}(\rho, \pi) = \begin{cases} \int \log(\frac{d\rho}{d\pi}) d\rho, & \text{when } \rho \ll \pi, \\ +\infty, & \text{otherwise.} \end{cases}$$

The following lemma shows in which sense the Kullback divergence function can be thought of as the dual of the log Laplace transform.

LEMMA 1.3. *For any bounded measurable function $h : \Theta \to \mathbb{R}$, and any probability distribution $\rho \in \mathcal{M}_+^1(\Theta)$ such that $\mathcal{K}(\rho, \pi) < \infty$,*

$$\log\bigl\{\pi\bigl[\exp(h)\bigr]\bigr\} = \rho(h) - \mathcal{K}(\rho, \pi) + \mathcal{K}(\rho, \pi_{\exp(h)}),$$

where by definition $\dfrac{d\pi_{\exp(h)}}{d\pi} = \dfrac{\exp[h(\theta)]}{\pi[\exp(h)]}$. *Consequently*

$$\log\bigl\{\pi\bigl[\exp(h)\bigr]\bigr\} = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho(h) - \mathcal{K}(\rho, \pi).$$

The proof is just a matter of writing down the definition of the quantities involved and using the fact that the Kullback divergence function is non negative. It can be found in [17, page 160]. In the duality between measurable functions and probability measures, we thus see that the log Laplace transform with respect to π is the Legendre transform of the Kullback divergence function with respect to π. Using this, we get

$$\mathbb{P}\Bigl\{\exp\bigl\{\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho[U_\lambda(\theta)] - \mathcal{K}(\rho, \pi)\bigr\}\Bigr\} \leq 1,$$

which, combined with the convexity of $\lambda \Phi_{\frac{\lambda}{N}}$, proves the basic inequality we were looking for.

THEOREM 1.4. *For any real constant λ,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\Bigl[\rho\bigl(\Phi_{\frac{\lambda}{N}} \circ R\bigr) - \rho(r)\Bigr] - \mathcal{K}(\rho, \pi)\biggr]\biggr\}$$
$$\leq \mathbb{P}\biggl\{\exp\biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\Bigl[\Phi_{\frac{\lambda}{N}}\bigl[\rho(R)\bigr] - \rho(r)\Bigr] - \mathcal{K}(\rho, \pi)\biggr]\biggr\} \leq 1.$$

The following sections will show how to use this theorem.

1.2. NON LOCAL BOUNDS. At least three sorts of bounds can be deduced from Theorem 1.4.

The most interesting ones to build estimators and tune parameters, as well as the first that have been considered in the development of the PAC-Bayesian approach, are deviation bounds. They provide an empirical upper bound for $\rho(R)$ — that is a bound which can be computed from observed data — with some probability $1 - \epsilon$, where ϵ is a presumably small and tunable confidence level.

Anyhow, since most of the results about the convergence speed of estimators to be found in the statistical literature are concerned with the expectation $\mathbb{P}\bigl[\rho(R)\bigr]$, it is also enlightening to bound this quantity. In order to know at which rate it may be approaching $\inf_\Theta R$, a non random upper bound is required, which will relate the average of the expected risk $\mathbb{P}\bigl[\rho(R)\bigr]$ with the properties of the contrast function $\theta \mapsto R(\theta)$.

Since the values of constants do matter a lot when a bound is to be used to select between various estimators using classification models of various complexities, a third kind of bound, related to the first, may be considered for the sake of its hopefully better constants: we will call them *unbiased empirical bounds*, to stress the fact that they provide some empirical quantity whose expectation under $\mathbb{P}$ can be proved to be an upper bound for $\mathbb{P}\bigl[\rho(R)\bigr]$, the average expected risk. The price to pay for these better constants is of course the lack of formal guarantee given by the bound : two random variables whose expectations are ordered in a certain way may very well be ordered in the reverse way with a large probability, so that basing the estimation of parameters or the selection of an estimator on some unbiased empirical bound is a hazardous business. Anyhow, since it is common practice to use the inequalities provided by mathematical statistical theory while replacing the proven constants with smaller values showing a better practical efficiency, considering unbiased empirical bounds akin to deviation bounds provides an indication about how much the constants may be decreased while not violating the theory too outrageously.

1.2.1. Unbiased empirical bounds. Let $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ be some fixed (and arbitrary) posterior distribution, describing some randomized estimator of θ. As we already mentioned, in these notes a posterior distribution will always be a regular conditional probability measure. By this we mean that

- for any $A \in \mathcal{T}$, the map $\omega \mapsto \rho(\omega, A) : \bigl(\Omega, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\bigr) \to \mathbb{R}_+$ is assumed to be measurable;
- for any $\omega \in \Omega$, the map $A \mapsto \rho(\omega, A) : \mathcal{T} \to \mathbb{R}_+$ is assumed to be a probability measure.

We will also assume without further notice that the σ-algebras we deal with are always countably generated. The technical implications of these assumptions are standard and discussed for instance in [17, pages 50-54] (where, among other things, a detailed proof of the decomposition of the Kullback Liebler divergence is given).

Let us restrict to the case when the constant λ is positive. We get from Theorem 1.4 that

$$\exp\Bigl[\lambda\Bigl\{\Phi_{\frac{\lambda}{N}}\Bigl[\mathbb{P}\bigl[\rho(R)\bigr]\Bigr]-\mathbb{P}\bigl[\rho(r)\bigr]\Bigr\}-\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]\Bigr]\leq 1, \tag{1.1}$$

where we have used the convexity of the exp function and of $\Phi_{\frac{\lambda}{N}}$. Since we have restricted our attention to positive values of the constant λ, Equation (1.1) can also be written

$$\mathbb{P}\bigl[\rho(R)\bigr]\leq\Phi_{\frac{\lambda}{N}}^{-1}\Bigl\{\mathbb{P}\bigl[\rho(r)+\lambda^{-1}\mathcal{K}(\rho,\pi)\bigr]\Bigr\},$$

leading to

Theorem 1.5. *For any posterior distribution $\rho:\Omega\rightarrow\mathcal{M}_+^1(\Theta)$, for any positive parameter λ,*

$$\begin{aligned}\mathbb{P}\bigl[\rho(R)\bigr]&\leq\frac{1-\exp\Bigl[-N^{-1}\mathbb{P}\bigl[\lambda\rho(r)+\mathcal{K}(\rho,\pi)\bigr]\Bigr]}{1-\exp(-\frac{\lambda}{N})}\\&\leq\mathbb{P}\biggl\{\frac{\lambda}{N\bigl[1-\exp(-\frac{\lambda}{N})\bigr]}\biggl[\rho(r)+\frac{\mathcal{K}(\rho,\pi)}{\lambda}\biggr]\biggr\}.\end{aligned}$$

The last inequality provides the *unbiased empirical upper bound* for $\rho(R)$ we were looking for, meaning that the expectation of $\frac{\lambda}{N\bigl[1-\exp(-\frac{\lambda}{N})\bigr]}\Bigl[\rho(r)+\frac{\mathcal{K}(\rho,\pi)}{\lambda}\Bigr]$ is larger than the expectation of $\rho(R)$. Let us notice that $1\leq\frac{\lambda}{N\bigl[1-\exp(-\frac{\lambda}{N})\bigr]}\leq\bigl[1-\frac{\lambda}{2N}\bigr]^{-1}$ and therefore that this coefficient is close to 1 when λ is significantly smaller than N.

If we are ready to believe in this bound (although this belief is not mathematically well founded, as we already mentioned), we can use it to optimize λ and to choose ρ. While the optimal choice of ρ when λ is fixed is to take it equal to $\pi_{\exp(-\lambda r)}$, a Gibbs posterior distribution, as it is sometimes called, we may for computational reasons be more interested in choosing ρ in some other class of posterior distributions.

For instance, our real interest may be to select some deterministic estimator from a family $\widehat{\theta}_m:\Omega\rightarrow\Theta_m$, $m\in M$, of possible ones, where Θ_m are measurable subsets of Θ and where M is an arbitrary (non necessarily countable) index set. We may for instance think of the case when

$\widehat{\theta}_m \in \arg\min_{\Theta_m} r$. We may slightly randomize the estimators to start with, considering for any $\theta \in \Theta_m$ and any $m \in M$,

$$\Delta_m(\theta) = \Bigl\{ \theta' \in \Theta_m : \bigl[f_{\theta'}(X_i) \bigr]_{i=1}^N = \bigl[f_\theta(X_i) \bigr]_{i=1}^N \Bigr\},$$

and defining ρ_m by the formula

$$\frac{d\rho_m}{d\pi}(\theta) = \frac{\mathbb{1}\bigl[\theta \in \Delta_m(\widehat{\theta}_m)\bigr]}{\pi\bigl[\Delta_m(\widehat{\theta}_m)\bigr]}.$$

Our posterior is minimizing $\mathcal{K}(\rho, \pi)$ among those whose support is restricted to the values of θ in Θ_m for which the classification rule f_θ is identical to the estimated one $f_{\widehat{\theta}_m}$ on the observed sample. Presumably, in many practical situations, $f_\theta(x)$ will be ρ_m almost surely identical to $f_{\widehat{\theta}_m}(x)$ when θ is drawn from ρ_m, for the vast majority of the values of $x \in \mathcal{X}$ and all the submodels Θ_m not plagued with too much overfitting (since this is by construction the case when $x \in \{X_i : i = 1, \dots, N\}$). Therefore replacing $\widehat{\theta}_m$ with ρ_m can be expected to be a minor change in many situations. This change by the way can be estimated in the (admittedly not so common) case when the distribution of the patterns $(X_i)_{i=1}^N$ is known. Indeed, introducing the pseudo distance

$$D(\theta, \theta') = \frac{1}{N} \sum_{i=1}^N \mathbb{P}\bigl[f_\theta(X_i) \neq f_{\theta'}(X_i) \bigr], \qquad \theta, \theta' \in \Theta, \tag{1.2}$$

one immediately sees that $R(\theta') \leq R(\theta) + D(\theta, \theta')$, for any $\theta, \theta' \in \Theta$, and therefore that

$$R(\widehat{\theta}_m) \leq \rho_m(R) + \rho_m\bigl[D(\cdot, \widehat{\theta}_m) \bigr].$$

Let us notice also that in the case where $\Theta_m \subset \mathbb{R}^{d_m}$, and R happens to be convex on $\Delta_m(\widehat{\theta}_m)$, then $\rho_m(R) \geq R\bigl[\int \theta \rho_m(d\theta)\bigr]$, and we can replace $\widehat{\theta}_m$ with $\widetilde{\theta}_m = \int \theta \rho_m(d\theta)$, and obtain bounds for $R(\widetilde{\theta}_m)$. This is not a very heavy assumption about R, in the case where we consider $\widehat{\theta}_m \in \arg\min_{\Theta_m} r$. Indeed, $\widehat{\theta}_m$, and therefore $\Delta_m(\widehat{\theta}_m)$, will be presumably close to $\arg\min_{\Theta_m} R$, and requiring a function to be convex in the neighboorhood of its minima is not a very strong assumption.

Since $r(\widehat{\theta}_m) = \rho_m(r)$, and $\mathcal{K}(\rho_m, \pi) = -\log\bigl\{\pi\bigl[\Delta_m(\widehat{\theta}_m)\bigr]\bigr\}$, our unbiased empirical upper bound in this context reads as

$$\frac{\lambda}{N\bigl[1 - \exp(-\frac{\lambda}{N})\bigr]} \left\{ r(\widehat{\theta}_m) - \frac{\log\bigl\{\pi\bigl[\Delta_m(\widehat{\theta}_m)\bigr]\bigr\}}{\lambda} \right\}.$$

Let us notice that we obtain a complexity factor $-\log\bigl\{\pi\bigl[\Delta_m(\widehat{\theta}_m)\bigr]\bigr\}$ which may be compared with the Vapnik Cervonenkis dimension. Indeed, in the case of binary classification, when using a classification model with VC dimension not greater than h_m, that is when any subset of $\mathcal{X}$ which can be split in any arbitrary way by some classification rule f_θ of the model Θ_m has at most h_m points, then

$$\bigl\{\Delta_m(\theta) : \theta \in \Theta_m\bigr\}$$

is a partition of Θ_m with at most $\bigl(\frac{eN}{h}\bigr)^h$ components. Therefore

$$\inf_{\theta \in \Theta_m} -\log\bigl\{\pi\bigl[\Delta_m(\theta)\bigr]\bigr\} \leq h_m \log\left(\frac{eN}{h_m}\right) - \log\bigl[\pi(\Theta_m)\bigr].$$

Thus, if the model and prior distribution are well suited to the classification task, in the sense that there is more "room" (where room is measured with π) between the two clusters defined by $\widehat{\theta}_m$ than between other partitions of the sample of patterns $(X_i)_{i=1}^N$, then we will have

$$-\log\bigl\{\pi\bigl[\Delta_m(\widehat{\theta})\bigr]\bigr\} \leq h_m \log\left(\frac{eN}{h_m}\right) - \log\bigl[\pi(\Theta_m)\bigr].$$

An optimal value $\widehat{m}$ may be selected so that

$$\widehat{m} \in \arg\min_{m \in M} \left\{ \inf_{\lambda \in \mathbb{R}_+} \frac{\lambda}{N\bigl[1 - \exp(-\frac{\lambda}{N})\bigr]} \left(r(\widehat{\theta}_m) - \frac{\log\bigl\{\pi\bigl[\Delta_m(\widehat{\theta}_m)\bigr]\bigr\}}{\lambda} \right) \right\}.$$

Since $\rho_{\widehat{m}}$ is still another posterior distribution, we can be sure that

$$\begin{aligned}\mathbb{P}\Bigl\{R(\widehat{\theta}_{\widehat{m}}) - \rho_{\widehat{m}}\bigl[D(\cdot, \widehat{\theta}_{\widehat{m}})\bigr]\Bigr\} &\leq \mathbb{P}\bigl[\rho_{\widehat{m}}(R)\bigr] \\ &\leq \inf_{\lambda \in \mathbb{R}_+} \mathbb{P}\left\{ \frac{\lambda}{N\bigl[1 - \exp(-\frac{\lambda}{N})\bigr]} \left(r(\widehat{\theta}_{\widehat{m}}) - \frac{\log\bigl\{\pi\bigl[\Delta_{\widehat{m}}(\widehat{\theta}_{\widehat{m}})\bigr]\bigr\}}{\lambda} \right) \right\}.\end{aligned}$$

(Taking the infimum in λ inside the expectation with respect to $\mathbb{P}$ would be possible at the price of some supplementary technicalities and a slight increase of the bound that we prefer to postpone to the discussion of deviation bounds, since they are the only ones to provide a rigorous mathematical foundation to the adaptive selection of estimators.)

1.2.2. Optimizing explicitly the exponential parameter λ. We would like to deal in this section with some technical issue we think helpful to the understanding of Theorem 1.5 (see page 13): namely to investigate how the upper bound it provides could be optimized, or at least approximately optimized, in λ. It turns out that this can be done quite explicitely.

So we will consider in this discussion the posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ to be fixed, and our aim will be to eliminate the constant λ from the bound by choosing its value in some nearly optimal way as a function of $\mathbb{P}\bigl[\rho(r)\bigr]$, the average of the empirical risk, and of $\mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr]$, which controls overfitting.

Let the bound be written as

$$\varphi(\lambda) = \bigl[1 - \exp(-\tfrac{\lambda}{N})\bigr]^{-1} \Bigl\{ 1 - \exp\Bigl[-\tfrac{\lambda}{N} \mathbb{P}\bigl[\rho(r)\bigr] - N^{-1} \mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr] \Bigr] \Bigr\}.$$

We see that

$$N \frac{\partial}{\partial \lambda} \log\bigl[\varphi(\lambda)\bigr] = \frac{\mathbb{P}\bigl[\rho(r)\bigr]}{\exp\Bigl[\tfrac{\lambda}{N}\mathbb{P}\bigl[\rho(r)\bigr] + N^{-1}\mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr]\Bigr] - 1} - \frac{1}{\exp(\tfrac{\lambda}{N}) - 1}.$$

Thus, the optimal value for λ is such that

$$\bigl[\exp(\tfrac{\lambda}{N}) - 1\bigr] \mathbb{P}\bigl[\rho(r)\bigr] = \exp\Bigl[\tfrac{\lambda}{N}\mathbb{P}\bigl[\rho(r)\bigr] + N^{-1}\mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr]\Bigr] - 1.$$

Assuming that $1 \gg \frac{\lambda}{N}\mathbb{P}\bigl[\rho(r)\bigr] \gg \frac{\mathbb{P}[\mathcal{K}(\rho,\pi)]}{N}$, and keeping only higher order terms, we are led to choose

$$\lambda = \sqrt{\frac{2N\mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr]}{\mathbb{P}\bigl[\rho(r)\bigr]\bigl\{1 - \mathbb{P}\bigl[\rho(r)\bigr]\bigr\}}},$$

obtaining

Theorem 1.6. *For any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\bigl[\rho(R)\bigr] \leq \frac{1 - \exp\biggl\{ -\sqrt{\frac{2\mathbb{P}[\mathcal{K}(\rho,\pi)]\mathbb{P}[\rho(r)]}{N\{1 - \mathbb{P}[\rho(r)]\}}} - \frac{\mathbb{P}[\mathcal{K}(\rho,\pi)]}{N} \biggr\}}{1 - \exp\biggl\{ -\sqrt{\frac{2\mathbb{P}[\mathcal{K}(\rho,\pi)]}{N\mathbb{P}[\rho(r)]\{1 - \mathbb{P}[\rho(r)]\}}} \biggr\}}.$$

This result of course is not very useful in itself, since none of the two quantities $\mathbb{P}\bigl[\rho(r)\bigr]$ and $\mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr]$ are easy to evaluate. Anyhow it gives a hint that replacing them boldly with $\rho(r)$ and $\mathcal{K}(\rho, \pi)$ could produce something

close to a legitimate empirical upper bound for $\rho(R)$. We will see in the subsection about deviation bounds that this is indeed essentially true.

Let us remark that in the second section of these notes, we will see another way of bounding

$$\inf_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1} \left(q + \frac{d}{\lambda} \right), \text{ leading to}$$

THEOREM 1.7. *For any prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\bigl[\rho(R)\bigr] \leq \left(1 + \frac{2\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]}{N} \right)^{-1} \Biggl\{ \mathbb{P}\bigl[\rho(r)\bigr] + \frac{\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]}{N} \\ + \sqrt{\frac{2\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]\mathbb{P}\bigl[\rho(r)\bigr]\bigl\{1 - \mathbb{P}\bigl[\rho(r)\bigr]\bigr\}}{N} + \frac{\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]^2}{N^2}} \Biggr\},$$

$$\textit{as soon as } \mathbb{P}\bigl[\rho(r)\bigr] + \sqrt{\frac{\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]}{2N}} \leq \frac{1}{2},$$

$$\textit{and } \mathbb{P}\bigl[\rho(R)\bigr] \leq \mathbb{P}\bigl[\rho(r)\bigr] + \sqrt{\frac{\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]}{2N}} \textit{ otherwise.}$$

This theorem enlightens the influence of three terms on the average expected risk :

• the average empirical risk, $\mathbb{P}\bigl[\rho(r)\bigr]$, which as a rule will decrease as the size of the classification model increases, acts as a *bias* term, grasping the ability of the model to account for the observed sample itself;

• a *variance* term $\mathbb{P}\bigl[\rho(r)\bigr]\bigl\{1-\mathbb{P}\bigl[\rho(r)\bigr]\bigr\}$ is due to the random fluctuations of $\rho(r)$;

• a *complexity* term $\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]$, which as a rule will increase with the size of the classification model, eventually acts as a multiplier of the variance term.

We observed numerically that the bound provided by Theorem 1.6 is better than the more classical Vapnik's like bound of Theorem 1.7. For instance, when $N = 1000$, $\mathbb{P}\bigl[\rho(r)\bigr] = 0.2$ and $\mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr] = 10$, Theorem 1.6 gives a bound lower than 0.2604, whereas the more classical Vapnik's like approximation of Theorem 1.7 gives a bound larger than 0.2622. Numerical simulations tend to suggest the two bounds are always ordered in the same way, although this could be a little teadious to prove mathematically.

1.2.3. Non random bounds. It is time now to come to less tentative results and see how far is the average expected error rate $\mathbb{P}\bigl[\rho(R)\bigr]$ from its best possible value $\inf_\Theta R$.

Let us notice first that

$$\lambda\rho(r) + \mathcal{K}(\rho,\pi) = \mathcal{K}(\rho,\pi_{\exp(-\lambda r)}) - \log\Bigl\{\pi\bigl[\exp(-\lambda r)\bigr]\Bigr\}.$$

Let us remark moreover that $r \mapsto \log\Bigl[\pi\bigl[\exp(-\lambda r)\bigr]\Bigr]$ is a convex functional, a property which can be used in the following way:

$$\begin{aligned}\mathbb{P}\Bigl\{\log\Bigl[\pi\bigl[\exp(-\lambda r)\bigr]\Bigr]\Bigr\} &= \mathbb{P}\Bigl\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)} -\lambda\rho(r) - \mathcal{K}(\rho,\pi)\Bigr\}\\ &\geq \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \mathbb{P}\Bigl\{-\lambda\rho(r) - \mathcal{K}(\rho,\pi)\Bigr\} = \sup_{\rho\in\mathcal{M}_+^1(\Theta)} -\lambda\rho(R) - \mathcal{K}(\rho,\pi)\\ &= \log\Bigl\{\pi\bigl[\exp(-\lambda R)\bigr]\Bigr\} = -\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta. \quad (1.3)\end{aligned}$$

These remarks applied to Theorem 1.5 lead to

Theorem 1.8. *For any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, for any positive parameter λ,*

$$\begin{aligned}\mathbb{P}\bigl[\rho(R)\bigr] &\leq \frac{1-\exp\Bigl\{-\frac{1}{N}\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta - \frac{1}{N}\mathbb{P}\bigl[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\bigr]\Bigr\}}{1-\exp(-\frac{\lambda}{N})}\\ &\leq \frac{1}{N\bigl[1-\exp(-\frac{\lambda}{N})\bigr]}\Bigl\{\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta + \mathbb{P}\bigl[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\bigr]\Bigr\}.\end{aligned}$$

This theorem is particularly well fitted for the case of the Gibbs posterior distribution $\rho = \pi_{\exp(-\lambda r)}$, where the entropy factor cancels and where $\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr]$ is shown to be bound to get close to $\inf_\Theta R$ when N goes to ∞, as soon as λ/N goes to 0 while λ goes to $+\infty$.

We can elaborate on Theorem 1.8 and define a notion of dimension of (Θ, R), with margin $\eta \geq 0$ putting

$$\begin{aligned}d_\eta(\Theta,R) &= \sup_{\beta\in\mathbb{R}_+} \beta\bigl[\pi_{\exp(-\beta R)}(R) - \operatorname{ess\,inf}_\pi R - \eta\bigr]\\ &\leq -\log\Bigl\{\pi\bigl[R \leq \operatorname{ess\,inf}_\pi R + \eta\bigr]\Bigr\}. \quad (1.4)\end{aligned}$$

This last inequality can be established by the chain of inequalities:

$$\beta \pi_{\exp(-\beta R)}(R) \leq \int_0^\beta \pi_{\exp(-\gamma R)}(R) d\gamma = -\log\Bigl\{\pi\bigl[\exp(-\beta R)\bigr]\Bigr\}$$
$$\leq \beta\Bigl(\operatorname*{ess\,inf}_\pi R + \eta\Bigr) - \log\Bigl[\pi\bigl(R \leq \operatorname*{ess\,inf}_\pi R + \eta\bigr)\Bigr],$$

where we have used successively the fact that $\lambda \mapsto \pi_{\exp(-\lambda R)}(R)$ is decreasing (because it is the derivative of the concave function $\lambda \mapsto -\log\bigl\{\pi\bigl[\exp(-\lambda R)\bigr]\bigr\}$) and the fact that the exponential function takes positive values.

In typical "parametric" situations $d_0(\Theta, R)$ will be finite, and in all circumstances $d_\eta(\Theta, R)$ will be finite for any $\eta > 0$ (this is a direct consequence of the definition of the essential infimum). Using this notion of dimension, we see that

$$\int_0^\lambda \pi_{\exp(-\beta R)}(R) d\beta \leq \lambda\bigl(\operatorname*{ess\,inf}_\pi R + \eta\bigr) + \int_0^\lambda \left[\frac{d_\eta}{\beta} \wedge (1 - \operatorname*{ess\,inf}_\pi R - \eta)\right] d\beta$$
$$= \lambda\bigl(\operatorname*{ess\,inf}_\pi R + \eta\bigr) + d_\eta(\Theta, R) \log\left[\frac{e\lambda}{d_\eta(\Theta, R)}\bigl(1 - \operatorname*{ess\,inf}_\pi R - \eta\bigr)\right].$$

This leads to

Corollary 1.9 *With the above notations, for any margin $\eta \in \mathbb{R}_+$, for any posterior distibution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\bigl[\rho(R)\bigr] \leq \inf_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1} \left[\operatorname*{ess\,inf}_\pi R + \eta + \frac{d_\eta}{\lambda} \log\left(\frac{e\lambda}{d_\eta}\right) + \frac{\mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr]\bigr\}}{\lambda}\right].$$

If one is wanting a posterior distribution with a small support, the theorem can also be applied to the case when ρ is obtained by truncating $\pi_{\exp(-\lambda r)}$ to some level set to reduce its support: let $\Theta_p = \{\theta \in \Theta : r(\theta) \leq p\}$, and let us define for any $q \in)0, 1)$ the level $p_q = \inf\{p : \pi_{\exp(-\lambda r)}(\Theta_p) \geq q\}$, let us then define ρ_q by its density

$$\frac{d\rho_q}{d\pi_{\exp(-\lambda r)}}(\theta) = \frac{\mathbb{1}(\theta \in \Theta_{p_q})}{\pi_{\exp(-\lambda r)}(\Theta_{p_q})},$$

then $\rho_0 = \pi_{\exp(-\lambda r)}$ and for any $q \in (0, 1($,

$$\mathbb{P}\bigl[\rho_q(R)\bigr] \leq \frac{1 - \exp\left\{-\frac{1}{N}\int_0^\lambda \pi_{\exp(-\beta R)}(R) d\beta - \frac{\log(q)}{N}\right\}}{1 - \exp(-\frac{\lambda}{N})}$$
$$\leq \frac{1}{N\bigl[1 - \exp(-\frac{\lambda}{N})\bigr]}\left\{\int_0^\lambda \pi_{\exp(-\beta R)}(R) d\beta - \log(q)\right\}.$$

1.2.4. Deviation bounds. They provide results holding under the distribution $\mathbb{P}$ of the sample with probability at least $1-\epsilon$, for any given confidence level, set by the choice of $\epsilon \in)0,1($. Using them is the only way to be quite (i.e. with probability $1-\epsilon$) sure to do the right thing, although this right thing may be overpessimistic, since deviation upper bounds are larger than corresponding non biased bounds.

Starting again from Theorem 1.4, and using Markov's inequality $\mathbb{P}\bigl[\exp(h) \geq 1\bigr] \leq \mathbb{P}\bigl[\exp(h)\bigr]$, we obtain

THEOREM 1.10. *For any positive parameter λ, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}\rho(R) &\leq \Phi_{\frac{\lambda}{N}}^{-1}\left\{\rho(r) + \frac{\mathcal{K}(\rho,\pi) - \log(\epsilon)}{\lambda}\right\} \\ &= \frac{1 - \exp\left\{-\frac{\lambda\rho(r)}{N} - \frac{\mathcal{K}(\rho,\pi)-\log(\epsilon)}{N}\right\}}{1-\exp\bigl(-\frac{\lambda}{N}\bigr)} \\ &\leq \frac{\lambda}{N\bigl[1-\exp\bigl(-\frac{\lambda}{N}\bigr)\bigr]}\left[\rho(r) + \frac{\mathcal{K}(\rho,\pi)-\log(\epsilon)}{\lambda}\right].\end{aligned}$$

We see that for a fixed value of the parameter λ, the upper bound is optimized when the posterior is chosen to be the Gibbs distribution $\rho = \pi_{\exp(-\lambda r)}$.

Moreover we would like to be entitled to optimize the bound in λ. Gaining the required uniformity in λ can be done in the following way. Let us notice first that values of λ less than 1 are not interesting (because they provide a bound larger than one, at least as soon as $\epsilon \leq \exp(-1)$). Let us consider some real parameter $\alpha > 1$, and the set $\Lambda = \{\alpha^k; k \in \mathbb{N}\}$. Let us put on this set the probability measure $\nu(\alpha^k) = [(k+1)(k+2)]^{-1}$. Applying the previous theorem to $\lambda = \alpha^k$ at confidence level $1 - \frac{\epsilon}{(k+1)(k+2)}$, and using a union bound, we see that with probability at least $1-\epsilon$, for any posterior distribution ρ,

$$\rho(R) \leq \inf_{\lambda' \in \Lambda} \Phi_{\frac{\lambda'}{N}}^{-1}\left\{\rho(r) + \frac{\mathcal{K}(\rho,\pi) - \log(\epsilon) + 2\log\left[\frac{\log(\alpha^2\lambda')}{\log(\alpha)}\right]}{\lambda'}\right\}.$$

Now we can remark that for any $\lambda \in (1, +\infty($, there is $\lambda' \in \Lambda$ such that $\alpha^{-1}\lambda \leq \lambda' \leq \lambda$. Moreover, for any $q \in (0,1)$, $\beta \mapsto \Phi_\beta^{-1}(q)$ is increasing on $\mathbb{R}_+$. Thus with probability at least $1-\epsilon$, for any posterior distribution ρ,

$$\rho(R) \leq \inf_{\lambda \in (1,\infty(} \Phi_{\frac{\lambda}{N}}^{-1}\left\{\rho(r) + \frac{\alpha}{\lambda}\left[\mathcal{K}(\rho,\pi) - \log(\epsilon) + 2\log\left(\frac{\log(\alpha^2\lambda)}{\log(\alpha)}\right)\right]\right\}$$

$$= \inf_{\lambda \in (1, \infty(} \frac{1 - \exp\left\{ -\frac{\lambda}{N}\rho(r) - \frac{\alpha}{N}\left[\mathcal{K}(\rho,\pi) - \log(\epsilon) + 2\log\left(\frac{\log(\alpha^2\lambda)}{\log(\alpha)}\right)\right]\right\}}{1 - \exp(-\frac{\lambda}{N})}.$$

Taking the approximately optimal value

$$\lambda = \sqrt{\frac{2N\alpha\left[\mathcal{K}(\rho,\pi) - \log(\epsilon)\right]}{\rho(r)[1-\rho(r)]}},$$

we obtain

THEOREM 1.11. *With probability $1-\epsilon$, for any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, putting $d(\rho,\epsilon) = \mathcal{K}(\rho,\pi) - \log(\epsilon)$,*

$$\rho(R) \leq \inf_{k \in \mathbb{N}} \frac{1 - \exp\left\{ -\dfrac{\alpha^k}{N}\rho(r) - \dfrac{1}{N}\Big[d(\rho,\epsilon) + \log\big[(k+1)(k+2)\big]\Big]\right\}}{1 - \exp\left(-\dfrac{\alpha^k}{N}\right)}$$

$$\leq \frac{1 - \exp\left\{ -\sqrt{\dfrac{2\alpha\rho(r)d(\rho,\epsilon)}{N[1-\rho(r)]}} - \dfrac{\alpha}{N}\left[d(\rho,\epsilon) + 2\log\left(\frac{\log\left(\alpha^2\sqrt{\frac{2N\alpha d(\rho,\epsilon)}{\rho(r)[1-\rho(r)]}}\right)}{\log(\alpha)}\right)\right]\right\}}{1 - \exp\left[-\sqrt{\dfrac{2\alpha d(\rho,\epsilon)}{N\rho(r)[1-\rho(r)]}}\right]}.$$

Moreover with probability at least $1-\epsilon$, for any posterior distribution ρ such that $\rho(r) = 0$,

$$\rho(R) \leq 1 - \exp\left[-\frac{\mathcal{K}(\rho,\pi) - \log(\epsilon)}{N}\right].$$

We can also elaborate on the results in an other direction by introducing the *empirical dimension*

$$d_e = \sup_{\beta \in \mathbb{R}_+} \beta\big[\pi_{\exp(-\beta r)}(r) - \operatorname{ess\,inf}_{\pi} r\big] \leq -\log\big[\pi\big(r = \operatorname{ess\,inf}_{\pi} r\big)\big]. \qquad (1.5)$$

(There is no need to introduce a margin in this definition, since r takes at most N values, and therefore $\pi\big(r = \operatorname{ess\,inf}_{\pi} r\big)$ will be strictly positive.) This leads to

COROLLARY 1.12. *For any positive real constant λ, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\rho(R) \leq \Phi_{\frac{\lambda}{N}}^{-1}\left[\operatorname{ess\,inf}_{\pi} r + \frac{d_e}{\lambda}\log\left(\frac{e\lambda}{d_e}\right) + \frac{\mathcal{K}\big[\rho, \pi_{\exp(-\lambda r)}\big] - \log(\epsilon)}{\lambda}\right].$$

We could then make the bound uniform in λ and optimize this parameter in a way similar to what was done to obtain Theorem 1.11.

1.3. LOCAL BOUNDS. In this subsection, better bounds will be achieved through a better choice of the prior distribution. This better prior distribution turns out to depend on the unknown sample distribution $\mathbb{P}$, and some work is required to circumvent this and obtain empirical bounds.

1.3.1. Choice of the prior. As mentioned in the introduction, if one is willing to minimize the bound in expectation provided by Theorem 1.5 (page 13), one is led to consider the optimal choice $\pi = \mathbb{P}(\rho)$. However, this is but an ideal choice, since $\mathbb{P}$ is in all conceivable situations unknown. Nevertheless it shows that it is possible through Theorem 1.5 to measure the *complexity* of the classification model with $\mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \mathbb{P}(\rho)\bigr]\bigr\}$, which is nothing but the *mutual information* between the random sample $(X_i, Y_i)_{i=1}^N$ and the estimated parameter $\hat{\theta}$, when the sample is drawn according to $\mathbb{P}$ and the estimated parameter knowing the sample is drawn according to ρ.

In practice, since we cannot choose $\pi = \mathbb{P}(\rho)$, we have to be content with a *flat* prior π, resulting in a bound measuring complexity according to $\mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr] = \mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \mathbb{P}(\rho)\bigr]\bigr\} + \mathcal{K}\bigl[\mathbb{P}(\rho), \pi\bigr]$ larger by the entropy factor $\mathcal{K}\bigl[\mathbb{P}(\rho), \pi\bigr]$ than the optimal one (we are still commenting on Theorem 1.5).

If we want to base the choice of π on Theorem 1.8 (page 18), and if we choose $\rho = \pi_{\exp(-\lambda r)}$ to optimize this bound, we will be inclined to choose some π such that

$$\frac{1}{\lambda}\int_0^{\lambda} \pi_{\exp(-\beta R)}(R) d\beta = -\frac{1}{\lambda}\log\Bigl\{\pi\bigl[\exp(-\lambda R)\bigr]\Bigr\}$$

is as far as possible close to $\inf_{\theta \in \Theta} R(\theta)$ in all circumstances. To give some more specific example, in the case when the distribution of the design $(X_i)_{i=1}^N$ is known, one can introduce on the parameter space Θ the metric D already defined by equation (1.2, page 14) (or some available upper bound for this distance). In view of the fact that $R(\theta) - R(\theta') \leq D(\theta, \theta')$, for any $\theta, \theta' \in \Theta$, it can be meaningful, at least theoretically, to choose π as

$$\pi = \sum_{k=1}^{\infty} \frac{1}{k(k+1)} \pi_k,$$

where π_k is the uniform measure on some minimal (or close to minimal) 2^{-k}-net $\mathcal{N}(\Theta, D, 2^{-k})$ of the metric space (Θ, D). With this choice

$$- \frac{1}{\lambda} \log\Bigl\{\pi\bigl[\exp(-\lambda R)\bigr]\Bigr\} \leq \inf_{\theta \in \Theta} R(\theta) \\ + \inf_{k} \left\{ 2^{-k} + \frac{\log(|\mathcal{N}(\Theta, D, 2^{-k})|) + \log[k(k+1)]}{\lambda} \right\}.$$

Another possibility, when we have to deal with real valued parameters, meaning that $\Theta \subset \mathbb{R}^d$, is to code each real component $\theta_i \in \mathbb{R}$ of $\theta = (\theta_i)_{i=1}^d$ to some precision and to use a prior μ which is atomic on dyadic numbers. More precisely let us parametrize the set of dyadic real numbers as

$$\mathcal{D} = \Biggl\{ r\bigl[s, m, p, (b_j)_{j=1}^p\bigr] = s 2^m \Biggl(1 + \sum_{j=1}^{p} b_j 2^{-j} \Biggr) \\ : s \in \{-1, +1\}, m \in \mathbb{Z}, p \in \mathbb{N}, b_j \in \{0, 1\} \Biggr\},$$

where, as can be seen, s codes the sign, m the order of magnitude, p the precision and $(b_j)_{j=1}^p$ the binary representation of the dyadic number $r\bigl[s, m, p, (b_j)_{j=1}^p\bigr]$. We can for instance consider on $\mathcal{D}$ the probability distribution

$$\mu\bigl\{r\bigl[s, m, p, (b_j)_{j=1}^p\bigr]\bigr\} = \Bigl[3(|m|+1)(|m|+2)(p+1)(p+2)2^p\Bigr]^{-1}, \qquad (1.6)$$

and define $\pi \in \mathcal{M}_+^1(\mathbb{R}^d)$ as $\pi = \mu^{\otimes d}$. This kind of "coding" prior distribution can be used also to define a prior on the integers (by renormalizing the restriction of μ to integers to get a probability distribution). Using μ is somehow equivalent to picking up a representative of each dyadic interval, and makes it possible to restrict to the case when the posterior ρ is a Dirac mass without losing too much (when $\Theta = (0, 1)$, this approach is somewhat equivalent to considering as prior distribution the Lebesgue measure and using as posterior distributions the uniform probability measures on dyadic intervals, with the advantage of obtaining non randomized estimators). When one uses in this way an atomic prior and Dirac masses as posterior distributions, the bounds proven so far can be obtained through a simpler union bound argument. This is so true that some of the detractors of the PAC-Bayesian approach (which, as a newcomer, has sometimes received a suspicious greeting among statisticians) have argued that it cannot bring anything that elementary union bound arguments could not essentially provide. We do not share of course this derogatory opinion, and while we think that allowing for non atomic priors and posteriors is worthwhile, we

also would like to stress that next to come local and relative bounds could hardly be obtained with the only help of union bounds.

Although the choice of a *flat* prior seems at first glance to be the only alternative when nothing is known about the sample distribution $\mathbb{P}$, the previous discussion shows that this type of choice is lacking proper localisation, and namely that we loose a factor $\mathcal{K}\bigl\{\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}\bigr], \pi\bigr\}$, the divergence between the bound-optimal prior $\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}\bigr]$, which is concentrated near the minima of R in favourable situations, and the flat prior π. Fortunately, there are technical ways to get around this difficulty and to obtain more local empirical bounds.

1.3.2. Unbiased local empirical bounds. The idea is to start with some flat prior $\pi \in \mathcal{M}_+^1(\Theta)$, and the posterior distribution $\rho = \pi_{\exp(-\lambda r)}$ minimizing the bound of Theorem 1.5 (page 13), when π is used as a prior. To improve the bound, we would like to use $\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}\bigr]$ instead of π, and we are going to make the guess that we could approximate it with $\pi_{\exp(-\beta R)}$ (we have replaced the parameter λ with some distinct parameter β to give some more freedom to our investigation, and also because, intuitively, $\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}\bigr]$ may be expected to be less concentrated than each of the $\pi_{\exp(-\lambda r)}$ it is mixing, which suggests that the best approximation of $\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}\bigr]$ by some $\pi_{\exp(-\beta R)}$ may be obtained for some parameter $\beta < \lambda$). We are then led to look for some empirical upper bound of $\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr]$. This is happily provided by the following computation

$$\begin{aligned}\mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr]\bigr\} &= \mathbb{P}\bigl[\mathcal{K}(\rho, \pi)\bigr] + \beta \mathbb{P}\bigl[\rho(R)\bigr] + \log\Bigl\{\pi\bigl[\exp(-\beta R)\bigr]\Bigr\} \\ &= \mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} + \beta \mathbb{P}\bigl[\rho(R - r)\bigr] \\ &\qquad + \log\Bigl\{\pi\bigl[\exp(-\beta R)\bigr]\Bigr\} - \mathbb{P}\Bigl\{\log \pi\bigl[\exp(-\beta r)\bigr]\Bigr\}.\end{aligned}$$

Using the convexity of $r \mapsto \log\bigl\{\pi\bigl[\exp(-\beta r)\bigr]\bigr\}$ as in equation (1.3) on page 18, we see that

$$0 \leq \mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr]\bigr\} \leq \beta \mathbb{P}\bigl[\rho(R - r)\bigr] + \mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\}.$$

This inequality has an interest of its own, since it provides a lower bound for $\mathbb{P}\bigl[\rho(R)\bigr]$. Moreover we can plug it into Theorem 1.5 (page 13) applied to the prior distribution $\pi_{\exp(-\beta R)}$ and obtain for any posterior distribution ρ and any positive paramter λ that

$$\Phi_{\frac{\lambda}{N}}\bigl\{\mathbb{P}\bigl[\rho(R)\bigr]\bigr\} \leq \mathbb{P}\biggl\{\rho(r) + \frac{\beta}{\lambda}\rho(R - r) + \frac{1}{\lambda}\mathbb{P}\Bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\Bigr\}\biggr\}.$$

In view of this, it it convenient to introduce the function

$$\begin{aligned}\widetilde{\Phi}_{a,b}(p) &= (1-b)^{-1}\bigl[\Phi_a(p) - bp\bigr] \\ &= -(1-b)^{-1}\Bigl\{a^{-1}\log\bigl\{1 - p\bigl[1-\exp(-a)\bigr]\bigr\} + bp\Bigr\}, \\ &\qquad\qquad p \in (0,1), a \in)0,\infty(, b \in (0,1(.\end{aligned}$$

This is a convex function of p, moreover

$$\widetilde{\Phi}'_{a,b}(0) = \Bigl\{a^{-1}\bigl[1-\exp(-a)\bigr] - b\Bigr\}(1-b)^{-1},$$

showing that it is an increasing one to one convex map of the unit interval unto itself as soon as $b \leq a^{-1}\bigl[1-\exp(-a)\bigr]$. Its convexity, combined with the value of its derivative at the origin, shows that

$$\widetilde{\Phi}_{a,b}(p) \geq \frac{a^{-1}\bigl[1-\exp(-a)\bigr] - b}{1-b} p.$$

Using these notations and remarks, we can state

THEOREM 1.13. *For any positive real constants β and λ such that $0 \leq \beta < N[1-\exp(-\frac{\lambda}{N})]$, for any posterior distribution $\rho : \Omega \to \mathcal{M}^1_+(\Theta)$,*

$$\begin{aligned}\mathbb{P}\biggl\{\rho(r) - \frac{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]}{\beta}\biggr\} &\leq \mathbb{P}\bigl[\rho(R)\bigr] \\ &\leq \widetilde{\Phi}^{-1}_{\frac{\lambda}{N},\frac{\beta}{\lambda}}\biggl\{\mathbb{P}\biggl[\rho(r) + \frac{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]}{\lambda-\beta}\biggr]\biggr\} \\ &\leq \frac{\lambda-\beta}{N[1-\exp(-\frac{\lambda}{N})] - \beta}\mathbb{P}\biggl[\rho(r) + \frac{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]}{\lambda-\beta}\biggr].\end{aligned}$$

Thus (taking $\lambda = 2\beta$), for any β such that $0 \leq \beta < \frac{N}{2}$,

$$\mathbb{P}\bigl[\rho(R)\bigr] \leq \frac{1}{1-\frac{2\beta}{N}}\mathbb{P}\biggl\{\rho(r) + \frac{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]}{\beta}\biggr\}.$$

Note that the last inequality is obtained using the fact that $1-\exp(-x) \geq x - \frac{x^2}{2}$, $x \in \mathbb{R}_+$.

COROLLARY 1.14. *For any $\beta \in (0, N($,*

$$\mathbb{P}\bigl[\pi_{\exp(-\beta r)}(r)\bigr] \leq \mathbb{P}\bigl[\pi_{\exp(-\beta r)}(R)\bigr] \\ \leq \inf_{\lambda \in (-N\log(1-\frac{\beta}{N}),\infty(} \frac{\lambda-\beta}{N[1-\exp(-\frac{\lambda}{N})]-\beta} \mathbb{P}\bigl[\pi_{\exp(-\beta r)}(r)\bigr] \\ \leq \frac{1}{1-\frac{2\beta}{N}} \mathbb{P}\bigl[\pi_{\exp(-\beta r)}(r)\bigr],$$

the last inequality holding only when $\beta < \frac{N}{2}$.

It is interesting to compare the upper bound provided by this corollary with Theorem 1.5 on page 13 when the posterior is a Gibbs measure $\rho = \pi_{\exp(-\beta r)}$. We see that we have succeeded to get rid of the entropy term $\mathcal{K}\bigl[\pi_{\exp(-\beta r)}, \pi\bigr]$, but at the price of an increase of the multiplicative factor, which for small values of $\frac{\beta}{N}$ grows from $(1-\frac{\beta}{2N})^{-1}$ (when we take $\lambda = \beta$ in Theorem 1.5), to $(1-\frac{2\beta}{N})^{-1}$. Therefore non localized bounds have an interest of their own, and are superseded by localized bounds only in favourable circumstances (presumably when the sample is large enough when compared with the complexity of the classification model).

Corollary 1.14 shows that when $\frac{2\beta}{N}$ is small, $\pi_{\exp(-\beta r)}(r)$ is a tight approximation of $\pi_{\exp(-\beta r)}(R)$ in the mean (since we have an upper bound and a lower bound which are close together).

Another corollary is obtained by optimizing the bound given by Theorem 1.13 in ρ, which is done by taking $\rho = \pi_{\exp(-\lambda r)}$.

COROLLARY 1.15. *For any positive real constants* β *and* λ *such that* $0 \leq \beta < N[1-\exp(-\frac{\lambda}{N})]$,

$$\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr] \leq \widetilde{\Phi}^{-1}_{\frac{\lambda}{N},\frac{\beta}{\lambda}}\biggl\{\mathbb{P}\biggl[\frac{1}{\lambda-\beta}\int_\beta^\lambda \pi_{\exp(-\gamma r)}(r)d\gamma\biggr]\biggr\} \\ \leq \frac{1}{N[1-\exp(-\frac{\lambda}{N})]-\beta}\mathbb{P}\biggl[\int_\beta^\lambda \pi_{\exp(-\gamma r)}(r)d\gamma\biggr].$$

Although this inequality gives by construction a better upper bound for $\inf_{\lambda\in\mathbb{R}_+}\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr]$ than Corollary 1.14, it is not easy to tell which one of the two inequalities is the best to bound $\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr]$ for a fixed (and possibly suboptimal) value of λ, because in this case, one factor is improved while the other is worsened.

Using the *empirical dimension* d_e defined by equation (1.5) on page 21, we see that

$$\frac{1}{\lambda-\beta}\int_\beta^\lambda \pi_{\exp(-\gamma r)}(r)d\gamma \leq \operatorname{ess\,inf}_\pi r + d_e \log\left(\frac{\lambda}{\beta}\right).$$

Therefore, in the case when we keep the ratio $\frac{\lambda}{\beta}$ bounded, we get a better dependence on the empirical dimension d_e than in Corollary 1.12 (page 21).

1.3.3. Non random local bounds. Let us come now to the localization of the non random upper bound given by Theorem 1.8 on page 18. According to Theorem 1.5 (page 13) applied to the localized prior $\pi_{\exp(-\beta R)}$,

$$\begin{aligned}
&\lambda \Phi_{\frac{\lambda}{N}}\bigl\{\mathbb{P}\bigl[\rho(R)\bigr]\bigr\} \leq \mathbb{P}\Bigl\{\lambda\rho(r) + \mathcal{K}(\rho,\pi) + \beta\rho(R)\Bigr\} + \log\bigl\{\pi\bigl[\exp(-\beta R)\bigr]\bigr\} \\
&= \mathbb{P}\Bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr] - \log\bigl\{\pi\bigl[\exp(-\lambda r)\bigr]\bigr\} + \beta\rho(R)\Bigr\} + \log\bigl\{\pi\bigl[\exp(-\beta R)\bigr]\bigr\} \\
&\leq \mathbb{P}\Bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr] + \beta\rho(R)\Bigr\} - \log\bigl\{\pi\bigl[\exp(-\lambda R)\bigr]\bigr\} + \log\bigl\{\pi\bigl[\exp(-\beta R)\bigr]\bigr\},
\end{aligned}$$

where we have used as previously inequality (1.3) (page 18). This proves

Theorem 1.16. *For any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, for any real parameters β and λ such that $0 \leq \beta < N\bigl[1 - \exp(-\frac{\lambda}{N})\bigr]$,*

$$\begin{aligned}
\mathbb{P}\bigl[\rho(R)\bigr] &\leq \widetilde{\Phi}^{-1}_{\frac{\lambda}{N},\frac{\beta}{\lambda}}\biggl\{\frac{1}{\lambda-\beta}\int_\beta^\lambda \pi_{\exp(-\gamma R)}(R)d\gamma + \mathbb{P}\biggl[\frac{\mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr]}{\lambda-\beta}\biggr]\biggr\} \\
&\leq \frac{1}{N\bigl[1-\exp(-\frac{\lambda}{N})\bigr]-\beta}\biggl\{\int_\beta^\lambda \pi_{\exp(-\gamma R)}(R)d\gamma + \mathbb{P}\Bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr]\Bigr\}\biggr\}.
\end{aligned}$$

Let us notice in particular that this theorem contains Theorem 1.8 (page 18) which corresponds to the case $\beta = 0$. As a corollary, we see also, taking $\rho = \pi_{\exp(-\lambda r)}$ and $\lambda = 2\beta$, and noticing that $\gamma \mapsto \pi_{\exp(-\gamma R)}(R)$ is decreasing, that

$$\begin{aligned}
\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr] &\leq \inf_{\beta,\beta<N[1-\exp(-\frac{\lambda}{N})]} \frac{\beta}{N\bigl[1-\exp(-\frac{\lambda}{N})\bigr]-\beta}\pi_{\exp(-\beta R)}(R) \\
&\leq \frac{1}{1-\frac{\lambda}{N}}\pi_{\exp(-\frac{\lambda}{2}R)}(R).
\end{aligned}$$

We can use this inequality in conjunction with the notion of dimension with margin η introduced by equation (1.4) on page 18, to see that the Gibbs posterior achieves for a proper choice of λ and any margin parameter $\eta \geq 0$ (which can be chosen to be equal to zero in parametric situations)

$$\inf_\lambda \mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr] \leq \operatorname{ess\,inf}_\pi R + \eta + \frac{4d_\eta}{N}$$

$$+ 2\sqrt{\frac{2d_\eta\bigl(\operatorname{ess\,inf}_\pi R + \eta\bigr)}{N} + \frac{4d_\eta^2}{N^2}}. \quad (1.7)$$

Deviation bounds to come next will show that the optimal λ can be estimated from empirical data.

Let us propose a little numerical example as an illustration : assuming that $d_0 = 10$, $N = 1000$ and $\operatorname{ess\,inf}_\pi R = 0.2$, we obtain from equation (1.7) that $\inf_\lambda \mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr] \leq 0.373$.

1.3.4. Local deviation bounds. When it comes to deviation bounds, we will for technical reasons choose a slightly more involved change of prior distribution and apply Theorem 1.10 (page 20) to the prior $\pi_{\exp[-\beta\Phi_{-\frac{\beta}{N}}\circ R]}$. The advantage of tweaking R with the nonlinear function $\Phi_{-\frac{\beta}{N}}$ will appear in the search for an empirical upper bound of the local entropy term. Theorem 1.4 (page 11), used with the above mentioned local prior, shows that

$$\mathbb{P}\left\{\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\Bigl\{\rho\bigl(\Phi_{\frac{\lambda}{N}}\circ R\bigr) - \rho(r)\Bigr\} - \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta\Phi_{-\frac{\beta}{N}}\circ R)}\bigr]\right\} \leq 1. \quad (1.8)$$

Moreover

$$\begin{aligned}\mathcal{K}\bigl[\rho, \pi_{\exp[-\beta\Phi_{-\frac{\beta}{N}}\circ R]}\bigr] &= \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr] + \beta\rho\Bigl[\Phi_{-\frac{\beta}{N}}\circ R - r\Bigr] \\ &\quad + \log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr]\Bigr\} - \log\Bigl\{\pi\Bigl[\exp(-\beta r)\Bigr]\Bigr\}, \quad (1.9)\end{aligned}$$

which is an invitation to find an upper bound for $\log\Bigl\{\pi\Bigl[\exp\bigl[-\beta\Phi_{-\frac{\lambda}{N}}\circ R\bigr]\Bigr]\Bigr\} - \log\Bigl\{\pi\bigl[\exp(-\beta r)\bigr]\Bigr\}$. Let us call for short $\overline{\pi}$ our localized prior distribution, thus defined as

$$\frac{d\overline{\pi}}{d\pi}(\theta) = \frac{\exp\Bigl\{-\beta\Phi_{-\frac{\beta}{N}}\bigl[R(\theta)\bigr]\Bigr\}}{\pi\Bigl\{\exp\bigl[-\beta\Phi_{-\frac{\beta}{N}}\circ R\bigr]\Bigr\}}.$$

Applying once again Theorem 1.4 (page 11), but this time to $-\beta$, we see that

$$\begin{aligned}&\mathbb{P}\left\{\exp\Bigl[\log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr]\Bigr\} - \log\Bigl\{\pi\bigl[\exp(-\beta r)\bigr]\Bigr\}\Bigr]\right\} \\ &\quad = \mathbb{P}\left\{\exp\Bigl[\log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R)\bigr)\Bigr]\Bigr\} + \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \beta\rho(r) + \mathcal{K}(\rho, \pi)\Bigr]\right\}\end{aligned}$$

$$\leq \mathbb{P}\biggl\{\exp\biggl[\log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R)\bigr)\Bigr]\Bigr\} + \beta\overline{\pi}(r) + \mathcal{K}(\overline{\pi},\pi)\biggr]\biggr\}$$

$$= \mathbb{P}\biggl\{\exp\biggl[\beta\Bigl[\overline{\pi}(r) - \overline{\pi}\bigl(\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr] - \mathcal{K}(\overline{\pi},\overline{\pi})\biggr]\biggr\} \leq 1. \quad (1.10)$$

Combining equations (1.9) and (1.10) and using the concavity of $\Phi_{-\frac{\beta}{N}}$, we see that with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$0 \leq \mathcal{K}(\rho,\overline{\pi}) \leq \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr] + \beta\Bigl[\Phi_{-\frac{\beta}{N}}\bigl[\rho(R)\bigr] - \rho(r)\Bigr] - \log(\epsilon).$$

We have proved a lower deviation bound:

Theorem 1.17 *For any positive real constant β, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\frac{\exp\biggl\{\dfrac{\beta}{N}\biggl[\rho(r) - \dfrac{\mathcal{K}[\rho,\pi_{\exp(-\beta r)}] - \log(\epsilon)}{\beta}\biggr]\biggr\} - 1}{\exp\bigl(\frac{\beta}{N}\bigr) - 1} \leq \rho(R).$$

Let us now seek for an upper bound. Using the Cauchy-Schwarz inequality to combine equations (1.8) and (1.10), we obtain

$$\mathbb{P}\biggl\{\exp\biggl[\frac{1}{2}\sup_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda\rho\bigl(\Phi_{\frac{\lambda}{N}}\circ R\bigr) - \beta\rho\bigl(\Phi_{-\frac{\beta}{N}}\circ R\bigr) - (\lambda-\beta)\rho(r) - \mathcal{K}\bigl[\rho,\pi_{\exp(-\beta r)}\bigr]\biggr]\biggr\}$$

$$= \mathbb{P}\biggl\{\exp\biggl[\tfrac{1}{2}\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Bigl(\lambda\Bigl\{\rho\bigl(\Phi_{\frac{\lambda}{N}}\circ R\bigr) - \rho(r)\Bigr\} - \mathcal{K}(\rho,\overline{\pi})\Bigr)\biggr]$$

$$\times \exp\biggl[\tfrac{1}{2}\Bigl(\log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr]\Bigr\} - \log\Bigl\{\pi\Bigl[\exp(-\beta r)\Bigr]\Bigr\}\Bigr)\biggr]\biggr\}$$

$$\leq \mathbb{P}\biggl\{\exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Bigl(\lambda\Bigl\{\rho\bigl(\Phi_{\frac{\lambda}{N}}\circ R\bigr) - \rho(r)\Bigr\} - \mathcal{K}(\rho,\overline{\pi})\Bigr)\biggr]\biggr\}^{1/2}$$

$$\times \mathbb{P}\biggl\{\exp\biggl[\Bigl(\log\Bigl\{\pi\Bigl[\exp\bigl(-\beta\Phi_{-\frac{\beta}{N}}\circ R\bigr)\Bigr]\Bigr\} - \log\Bigl\{\pi\Bigl[\exp(-\beta r)\Bigr]\Bigr\}\Bigr)\biggr]\biggr\}^{1/2} \leq 1. \quad (1.11)$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution ρ,

$$\lambda\Phi_{\frac{\lambda}{N}}\bigl[\rho(R)\bigr] - \beta\Phi_{-\frac{\beta}{N}}\bigl[\rho(R)\bigr] \leq (\lambda-\beta)\rho(r) + \mathcal{K}(\rho,\pi_{\exp(-\beta r)}) - 2\log(\epsilon).$$

(It would have been more straightforward to use a union bound on deviation inequalities instead of the Cauchy-Schwarz inequality on exponential moments, anyhow, this would have led to replace $-2\log(\epsilon)$ with the worse factor $2\log(\frac{2}{\epsilon})$.) Let us now remind that

$$\lambda \Phi_{\frac{\lambda}{N}}(p) - \beta \Phi_{-\frac{\beta}{N}}(p) = -N \log\Bigl\{1 - \bigl[1 - \exp\bigl(-\tfrac{\lambda}{N}\bigr)\bigr]p\Bigr\} \\ - N \log\Bigl\{1 + \bigl[\exp\bigl(\tfrac{\beta}{N}\bigr) - 1\bigr]p\Bigr\},$$

and let us put

$$B = (\lambda - \beta)\rho(r) + \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr] - 2\log(\epsilon) \\ = \mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr] + \int_{\beta}^{\lambda} \pi_{\exp(-\xi r)}(r) d\xi - 2\log(\epsilon).$$

Let us consider moreover the change of variables $\alpha = 1 - \exp(-\frac{\lambda}{N})$ and $\gamma = \exp(\frac{\beta}{N}) - 1$.
We obtain $\bigl[1 - \alpha\rho(R)\bigr]\bigl[1 + \gamma\rho(R)\bigr] \geq \exp(-\frac{B}{N})$, leading to

THEOREM 1.18. *For any positive constants α, γ, such that $0 \leq \gamma < \alpha < 1$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, the bound*

$$M(\rho) = -\frac{\log\bigl[(1-\alpha)(1+\gamma)\bigr]}{\alpha - \gamma}\rho(r) + \frac{\mathcal{K}(\rho, \pi_{\exp[-N\log(1+\gamma)r]}) - 2\log(\epsilon)}{N(\alpha - \gamma)} \\ = \frac{\mathcal{K}\bigl[\rho, \pi_{\exp[N\log(1-\alpha)r]}\bigr] + \displaystyle\int_{N\log(1+\gamma)}^{-N\log(1-\alpha)} \pi_{\exp(-\xi r)}(r) d\xi - 2\log(\epsilon)}{N(\alpha - \gamma)},$$

is such that

$$\rho(R) \leq \frac{\alpha - \gamma}{2\alpha\gamma}\left(\sqrt{1 + \frac{4\alpha\gamma}{(\alpha-\gamma)^2}\bigl\{1 - \exp\bigl[-(\alpha-\gamma)M(\rho)\bigr]\bigr\}} - 1\right) \leq M(\rho),$$

Using the *empirical dimension* d_e defined by equation (1.5) on page 21, we can use the inequality

$$\int_{\beta}^{\lambda} \pi_{\exp(-\xi r)}(r) d\xi \leq (\lambda - \beta) \operatorname*{ess\,inf}_{\pi} r + d_e \log\left(\frac{\lambda}{\beta}\right),$$

to prove that

$$M(\rho) \leq \frac{\log\bigl[(1+\gamma)(1-\alpha)\bigr]}{\gamma-\alpha} \operatorname*{ess\,inf}_{\pi} r + \frac{d_e \log\left[\frac{-\log(1-\alpha)}{\log(1+\gamma)}\right] + \mathcal{K}\bigl[\rho, \pi_{\exp[N\log(1-\alpha)r]}\bigr] - 2\log(\epsilon)}{N(\alpha-\gamma)}.$$

Let us give a little numerical illustration : assuming that $d_e = 10$ and $N = 1000$, taking $\epsilon = 0.01$, $\alpha = 0.5$ and $\gamma = 0.1$, we obtain from Theorem 1.18 $\pi_{\exp[N\log(1-\alpha)r]}(R) \simeq \pi_{\exp(-693r)}(R) \leq 0.332 \leq 0.372$, where we have given respectively the non linear and the linear bound. This shows the practical interest of keeping the non-linearity. Let us also mention that optimizing the values of the parameters α and γ would not have yielded a significantly lower bound.

The following corollary is obtained by taking $\lambda = 2\beta$ and keeping only the linear bound, we give it for the sake of its simplicity:

COROLLARY 1.19. *For any positive real constant β such that* $\exp(\frac{\beta}{N}) + \exp(-\frac{2\beta}{N}) < 2$, *which is the case when $\beta < 0.48N$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\rho(R) \leq \frac{\beta\rho(r) + \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr] - 2\log(\epsilon)}{N\bigl[2 - \exp\bigl(\frac{\beta}{N}\bigr) - \exp\bigl(-\frac{2\beta}{N}\bigr)\bigr]} = \frac{\int_\beta^{2\beta} \pi_{\exp(-\xi r)}(r)d\xi + \mathcal{K}\bigl[\rho, \pi_{\exp(-2\beta r)}\bigr] - 2\log(\epsilon)}{N\bigl[2 - \exp(\frac{\beta}{N}) - \exp(-\frac{2\beta}{N})\bigr]}.$$

Let us mention that this corollary applied to the above numerical example gives $\pi_{\exp(-200r)}(R) \leq 0.475$ (when we take $\beta = 100$, consistently with the choice $\gamma = 0.1$).

1.3.5. Partially local bounds. Local bounds are suitable when the lowest values of the empirical error rate r are reached only on a small part of the parameter set Θ. When Θ is the disjoint union of submodels of different complexities, the minimum of r will as a rule not be "localized" in a way that calls for the use of local bounds. Just think for instance of the case when $\Theta = \bigsqcup_{m=1}^M \Theta_m$, where the sets $\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta_M$ are nested. In this case we will have $\inf_{\Theta_1} r \geq \inf_{\Theta_2} r \geq \cdots \geq \inf_{\Theta_M} r$, although Θ_M may be too large to be the right model to use. In this situation, we do not want to localize the bound completely. Let us make a more specific fancyful

but typical pseudo computation. Just imagine we have a countable collection $(\Theta_m)_{m \in M}$ of submodels. Let us assume we are interested in choosing between the estimators $\widehat{\theta}_m \in \arg\min_{\Theta_m} r$, maybe randomizing them (e.g. replacing them with $\pi^m_{\exp(-\lambda r)}$). Let us imagine moreover that we are in a typically parametric situation, where, for some priors $\pi^m \in \mathcal{M}_+^1(\Theta_m)$, $m \in M$, there is a "dimension" d_m such that $\lambda\bigl[\pi^m_{\exp(-\lambda r)}(r) - r(\widehat{\theta}_m)\bigr] \simeq d_m$. Let $\mu \in \mathcal{M}_+^1(M)$ be some distribution on the index set M. It is easy to see that $(\mu\pi)_{\exp(-\lambda r)}$ will typically not be properly local, in the sense that typically

$$
\begin{aligned}
(\mu\pi)_{\exp(-\lambda r)}(r) &= \frac{\mu\Bigl\{\pi_{\exp(-\lambda r)}(r)\pi\bigl[\exp(-\lambda r)\bigr]\Bigr\}}{\mu\Bigl\{\pi\bigl[\exp(-\lambda r)\bigr]\Bigr\}} \\
&\simeq \frac{\displaystyle\sum_{m \in M} \bigl[(\inf_{\Theta_m} r) + \tfrac{d_m}{\lambda}\bigr] \exp\bigl[-\lambda(\inf_{\Theta_m} r) - d_m \log\bigl(\tfrac{e\lambda}{d_m}\bigr)\bigr]\mu(m)}{\displaystyle\sum_{m \in M} \exp\Bigl[-\lambda(\inf_{\Theta_m} r) - d_m \log\bigl(\tfrac{e\lambda}{d_m}\bigr)\Bigr]\mu(m)} \\
&\simeq \biggl\{\inf_{m \in M} (\inf_{\Theta_m} r) + \tfrac{d_m}{\lambda} \log\bigl(\tfrac{e\lambda}{d_m \mu(m)}\bigr)\biggr\} \\
&\qquad + \log\biggl\{\sum_{m \in M} \exp\bigl[-d_m \log(\tfrac{\lambda}{d_m})\bigr]\mu(m)\biggr\}.
\end{aligned}
$$

where we have used the estimate

$$
\begin{aligned}
-\log\Bigl\{\pi\bigl[\exp(-\lambda r)\bigr]\Bigr\} &= \int_0^\lambda \pi_{\exp(-\beta r)}(r) d\beta \\
&\simeq \int_0^\lambda (\inf_{\Theta_m} r) + \bigl[\tfrac{d_m}{\beta} \wedge 1\bigr] d\beta \simeq \lambda(\inf_{\Theta_m} r) + d_m\bigl[\log\bigl(\tfrac{\lambda}{d_m}\bigr) + 1\bigr].
\end{aligned}
$$

Our approximations have no pretention to be rigorous or very accurate, but they nevertheless give the best order of magnitude we can expect in typical situations, and show that this order of magnitude is not what we are looking for: mixing different models with the help of μ spoils the localization, introducing a multiplier $\log\bigl(\frac{\lambda}{d_m}\bigr)$ to the dimension d_m which is precisely what we would have got if we had not localized at all the bound. What we would really like to do in such situations is to use a *partially localized* posterior distribution, such as $\mu^{\widehat{m}}_{\exp(-\lambda r)}$, where $\widehat{m}$ is an estimator of the best submodel to be used. While the most straightforward way to do this is to use a union bound on results obtained for each submodel Θ_m, we are going

here to show how to allow arbitrary posterior distributions on the index set (corresponding to a randomization of the choice of $\widehat{m}$).

Let us consider the framework we just mentioned: let the measurable parameter set $(\Theta, \mathcal{T})$ be a disjoint union of measurable submodels, $\Theta = \bigsqcup_{m \in M} \Theta_m$. Let the index set $(M, \mathcal{M})$ be some measurable space (most of the time it will be a countable set). Let $\mu \in \mathcal{M}_+^1(M)$ be a prior probability distribution on $(M, \mathcal{M})$. Let $\pi : M \rightarrow \mathcal{M}_+^1(\Theta)$ be a regular conditional probability measure such that $\pi(m, \Theta_m) = 1$, for any $m \in M$. Let $\mu\pi \in \mathcal{M}_+^1(M \times \Theta)$ be the product probability measure defined by $\mu\pi(h) = \int_{m \in M} \bigl(\int_{\theta \in \Theta} h(m, \theta) \pi(m, d\theta) \bigr) \mu(dm)$, for any bounded measurable function $h : M \times \Theta \rightarrow \mathbb{R}$. Let $\pi_{\exp(h)} \in \mathcal{M}_+(M \times \Theta)$ be the regular conditionnal probability measure defined by

$$\frac{d\pi_{\exp(h)}}{d\pi}(m, \theta) = \frac{\exp\bigl[h(\theta)\bigr]}{\pi\bigl[m, \exp(h)\bigr]},$$

where consistently with previous notations $\pi(m, h) = \int_{\Theta} h(m, \theta) \pi(m, d\theta)$ (we will also often use the less explicit notation $\pi(h)$). Let for short

$$U(\theta, \omega) = \lambda \Phi_{\frac{\lambda}{N}}\bigl[R(\theta)\bigr] - \beta \Phi_{-\frac{\beta}{N}}\bigl[R(\theta)\bigr] - (\lambda - \beta) r(\theta, \omega).$$

Integrating with respect to μ equation (1.11) on page 29, written in each submodel Θ_m using the prior distribution $\pi(m, \cdot)$, we see that

$$\begin{aligned}
&\mathbb{P}\Biggl\{\exp\Biggl[\sup_{\nu \in \mathcal{M}_+^1(M)} \sup_{\rho : M \rightarrow \mathcal{M}_+^1(\Theta)} \frac{1}{2}\Bigl[(\nu\rho)(U) - \nu\bigl\{\mathcal{K}\bigl(\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\}\Bigr] - \mathcal{K}(\nu, \mu)\Biggr]\Biggr\} \\
&\leq \mathbb{P}\Biggl\{\exp\Biggl[\sup_{\nu \in \mathcal{M}_+^1(M)} \frac{1}{2}\nu\Biggl(\sup_{\rho : M \rightarrow \mathcal{M}_+^1(\Theta)} \rho(U) - \mathcal{K}(\rho, \pi_{\exp(-\beta r)})\Biggr) - \mathcal{K}(\nu, \mu)\Biggr]\Biggr\} \\
&\qquad = \mathbb{P}\Biggl\{\mu\Biggl[\exp\Biggl\{\tfrac{1}{2}\sup_{\rho : M \rightarrow \mathcal{M}_+^1(\Theta)} \Bigl[\rho(U) - \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\Bigr]\Biggr\}\Biggr]\Biggr\} \\
&\qquad\quad = \mu\Biggl\{\mathbb{P}\Biggl[\exp\Biggl\{\tfrac{1}{2}\sup_{\rho : M \rightarrow \mathcal{M}_+^1(\Theta)} \Bigl[\rho(U) - \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\Bigr]\Biggr\}\Biggr]\Biggr\} \leq 1.
\end{aligned}$$

This proves that

$$\begin{aligned}
\mathbb{P}\Biggl\{\exp\Biggl[\frac{1}{2}\sup_{\nu \in \mathcal{M}_+^1(M)} \sup_{\rho : M \rightarrow \mathcal{M}_+^1(\Theta)} \lambda \Phi_{\frac{\lambda}{N}}\bigl[\nu\rho(R)\bigr] - \beta \Phi_{-\frac{\beta}{N}}\bigl[\nu\rho(R)\bigr] \\
- (\lambda - \beta)\nu\rho(r) - 2\mathcal{K}(\nu, \mu) - \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\}\Biggr]\Biggr\} \leq 1. \qquad (1.12)
\end{aligned}$$

Introducing the optimal value of r on each submodel $r^\star(m) = \operatorname{ess\,inf}_{\pi(m,\cdot)} r$ and the empirical dimensions

$$d_e(m) = \sup_{\xi \in \mathbb{R}_+} \xi \bigl[\pi_{\exp(-\xi r)}(m,r) - r^\star(m) \bigr],$$

we can thus state

Theorem 1.20. *For any positive real constants $\beta < \lambda$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$, for any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$\lambda \Phi_{\frac{\lambda}{N}} \bigl[\nu\rho(R) \bigr] - \beta \Phi_{-\frac{\beta}{N}} \bigl[\nu\rho(R) \bigr] \leq B_1(\nu,\rho),$$

$$\begin{aligned}
&\textit{where } B_1(\nu,\rho) = (\lambda-\beta)\nu\rho(r) + 2\mathcal{K}(\nu,\mu) + \nu\bigl\{ \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} - 2\log(\epsilon) \\
&= \nu\left[\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha \right] + 2\mathcal{K}(\nu,\mu) + \nu\bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr]\bigr\} - 2\log(\epsilon) \\
&\qquad = 2\log\left\{ \mu\left[\exp\left(-\frac{1}{2}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha \right)\right]\right\} \\
&\qquad\quad + 2\mathcal{K}\Bigl[\nu, \mu_{\left(\frac{\pi[\exp(-\lambda r)]}{\pi[\exp(-\beta r)]}\right)^{1/2}}\Bigr] + \nu\bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr]\bigr\} - 2\log(\epsilon), \\
&\textit{and therefore } B_1(\nu,\rho) \leq \nu\left[(\lambda-\beta) r^\star + \log\left(\tfrac{\lambda}{\beta}\right) d_e \right] + 2\mathcal{K}(\nu,\mu) \\
&\qquad\qquad + \nu\bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr]\bigr\} - 2\log(\epsilon), \\
&\textit{as well as } B_1(\nu,\rho) \leq 2\log\left\{ \mu\left[\exp\left(-\frac{1}{2} r^\star + \frac{1}{2}\log\left(\tfrac{\lambda}{\beta}\right) d_e \right)\right]\right\} \\
&\qquad\qquad + 2\mathcal{K}\Bigl[\nu, \mu_{\frac{\pi[\exp(-\lambda r)]}{\pi[\exp(-\beta r)]}}\Bigr] + \nu\bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr] - 2\log(\epsilon).
\end{aligned}$$

Thus, for any real constants α and γ such that $0 \leq \gamma < \alpha < 1$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$ and any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$, the bound

$$\begin{aligned}
B_2(\nu,\rho) &= -\frac{\log\bigl[(1-\alpha)(1+\gamma)\bigr]}{\alpha-\gamma}\nu\rho(r) + \frac{2\mathcal{K}(\nu,\mu) + \nu\Bigl\{\mathcal{K}\bigl[\rho,\pi_{(1+\gamma)^{-Nr}}\bigr]\Bigr\} - 2\log(\epsilon)}{N(\alpha-\gamma)} \\
&= \frac{2\mathcal{K}\Bigl[\nu,\mu_{\left(\frac{\pi[(1-\alpha)^{Nr}]}{\pi[(1+\gamma)^{-Nr}]}\right)^{1/2}}\Bigr] + \nu\Bigl\{\mathcal{K}\bigl[\rho,\pi_{(1-\alpha)^{Nr}}\bigr]\Bigr\}}{N(\alpha-\gamma)} \\
&\qquad - \frac{2\log\left\{\mu\left[\exp\left[-\frac{1}{2}\int_{N\log(1+\gamma)}^{-N\log(1-\alpha)} \pi_{\exp(-\xi r)}(\cdot,r) d\xi\right]\right]\right\} + 2\log(\epsilon)}{N(\alpha-\gamma)}
\end{aligned}$$

satisfies

$$\nu\rho(R) \leq \frac{\alpha-\gamma}{2\alpha\gamma}\left(\sqrt{1+\frac{4\alpha\gamma}{(\alpha-\gamma)^2}\Bigl\{1-\exp\bigl[-(\alpha-\gamma)B(\nu,\rho)\bigr]\Bigr\}}-1\right) \leq B(\nu,\rho).$$

Let us remark that in the case when $\nu = \mu_{\left(\frac{\pi[(1-\alpha)^{Nr}]}{\pi[(1+\gamma)^{-Nr}]}\right)^{1/2}}$ and $\rho = \pi_{(1-\alpha)^{Nr}}$, we get as desired a bound that is adaptively local in all the Θ_m (at least when M is countable and μ is atomic):

$$\begin{aligned} B(\nu,\rho) \leq -\tfrac{2}{N(\alpha-\gamma)}\log\Biggl\{\mu\Biggl\{\exp\Biggl[\tfrac{N}{2}\log\bigl[(1+\gamma)(1-\alpha)\bigr]r^\star \\ -\log\left(\tfrac{-\log(1-\alpha)}{\log(1+\gamma)}\right)\tfrac{d_e}{2}\Biggr]\Biggr\}\Biggr\} - \frac{2\log(\epsilon)}{N(\alpha-\gamma)} \\ \leq \inf_{m\in M}\Biggl\{-\tfrac{\log\bigl[(1-\alpha)(1+\gamma)\bigr]}{\alpha-\gamma}r^\star(m) \\ +\log\left(\tfrac{-\log(1-\alpha)}{\log(1+\gamma)}\right)\tfrac{d_e(m)}{N(\alpha-\gamma)} - 2\tfrac{\log\bigl[\epsilon\mu(m)\bigr]}{N(\alpha-\gamma)}\Biggr\}. \end{aligned}$$

The penalization by the *empirical dimension* $d_e(m)$ in each submodel is as desired linear in $d_e(m)$. Non random partially local bounds could be obtained in a way that is easy to imagine. We leave this investigation to the reader.

1.3.6. Two step localization. We have seen that the bound optimal choice of the posterior distribution ν on the index set in Theorem 1.20 (page 34) is such that

$$\frac{d\nu}{d\mu}(m) \sim \left(\frac{\pi\bigl[\exp\bigl(-\lambda r(m,\cdot)\bigr)\bigr]}{\pi\bigl[\exp\bigl(-\beta r(m,\cdot)\bigr)\bigr]}\right)^{\frac{1}{2}} = \exp\left[-\frac{1}{2}\int_\beta^\lambda \pi_{\exp(-\alpha r)}(m,r)d\alpha\right].$$

This suggests to replace the prior distribution μ with $\overline{\mu}$ defined by its density

$$\frac{d\overline{\mu}}{d\mu}(m) = \frac{\exp\bigl[-h(m)\bigr]}{\mu\bigl[\exp(-h)\bigr]},$$

$$\text{where } h(m) = -\xi\int_\beta^\gamma \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\bigl[\Phi_{-\frac{\eta}{N}}\circ R(m,\cdot)\bigr]d\alpha. \quad (1.13)$$

The use of $\Phi_{-\frac{\eta}{N}}\circ R$ instead of R is motivated by technical reasons which will appear in subsequent computations. Indeed, we will need to bound

$$\nu\left[\int_\beta^\lambda \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\bigl(\Phi_{-\frac{\eta}{N}}\circ R\bigr)d\alpha\right]$$

in order to handle $\mathcal{K}(\nu, \overline{\mu})$. In the spirit of equation (1.8, page 28), starting back from Theorem 1.4 (page 11), applied in each submodel Θ_m to the prior distribution $\pi_{\exp(-\gamma\Phi_{-\frac{\eta}{N}} \circ R)}$ and integrated with respect to $\overline{\mu}$, we see that for any positive real constants λ, γ and η, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \rightarrow \mathcal{M}_+^1(M)$ on the index set and any conditional posterior distribution $\rho : \Omega \times M \rightarrow \mathcal{M}_+^1(\Theta)$,

$$\nu\rho\bigl(\lambda\Phi_{\frac{\lambda}{N}} \circ R - \gamma\Phi_{-\frac{\eta}{N}} \circ R\bigr) \leq \lambda\nu\rho(r) \\ + \nu\mathcal{K}(\rho, \pi) + \mathcal{K}(\nu, \overline{\mu}) + \nu\Bigl\{\log\Bigl[\pi\bigl[\exp\bigl(-\gamma\Phi_{-\frac{\eta}{N}} \circ R\bigr)\bigr]\Bigr]\Bigr\} - \log(\epsilon). \tag{1.14}$$

Since $x \mapsto f(x) \overset{\text{def}}{=} \lambda\Phi_{\frac{\lambda}{N}} - \gamma\Phi_{-\frac{\eta}{N}}(x)$ is a convex function, it is such that

$$f(x) \geq xf'(0) = xN\Bigl\{\bigl[1 - \exp(-\tfrac{\lambda}{N})\bigr] + \tfrac{\gamma}{\eta}\bigl[\exp(\tfrac{\eta}{N}) - 1\bigr]\Bigr\}.$$

Thus if we put

$$\gamma = \frac{\eta\bigl[1 - \exp(-\frac{\lambda}{N})\bigr]}{\exp(\frac{\eta}{N}) - 1}, \tag{1.15}$$

we obtain that $f(x) \geq 0$, $x \in \mathbb{R}$, and therefore that the left-hand side of equation (1.14) is non negative. We can moreover introduce the prior conditional distribution $\overline{\pi}$ defined by

$$\frac{d\overline{\pi}}{d\pi}(m, \theta) = \frac{\exp\bigl[-\beta\Phi_{-\frac{\eta}{N}} \circ R(\theta)\bigr]}{\pi\bigl\{m, \exp\bigl[-\beta\Phi_{-\frac{\eta}{N}} \circ R\bigr]\bigr\}}.$$

With $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\nu\Omega \rightarrow \mathcal{M}_+^1(M)$ and $\rho : \Omega \times M \rightarrow \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}\beta\nu\rho(r) + \nu\bigl[\mathcal{K}(\rho, \pi)\bigr] &= \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} - \nu\Bigl[\log\Bigl\{\pi\bigl[\exp(-\beta r)\bigr]\Bigr\}\Bigr] \\ &\leq \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} + \beta\nu\overline{\pi}(r) + \nu\bigl[\mathcal{K}(\overline{\pi}, \pi)\bigr] \\ &\leq \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} + \beta\nu\overline{\pi}\bigl(\Phi_{-\frac{\eta}{N}} \circ R\bigr) \\ &\qquad + \tfrac{\beta}{\eta}\bigl[\mathcal{K}(\nu, \overline{\mu}) - \log(\epsilon)\bigr] + \nu\bigl[\mathcal{K}(\overline{\pi}, \pi)\bigr] \\ &= \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} - \nu\Bigl\{\log\Bigl[\pi\bigl[\exp\bigl(-\beta\Phi_{-\frac{\eta}{N}} \circ R\bigr)\bigr]\Bigr]\Bigr\} \\ &\qquad + \tfrac{\beta}{\eta}\bigl[\mathcal{K}(\nu, \overline{\mu}) - \log(\epsilon)\bigr].\end{aligned}$$

Thus, coming back to equation (1.14), we see that under condition (1.15), with $\mathbb{P}$ probability at least $1-\epsilon$,

$$0 \leq (\lambda - \beta)\nu\rho(r) + \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} \\ - \nu\biggl[\int_{\beta}^{\gamma} \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\bigl(\Phi_{-\frac{\eta}{N}}\circ R\bigr)d\alpha\biggr] + (1+\tfrac{\beta}{\eta})\bigl[\mathcal{K}(\nu,\overline{\mu}) + \log(\tfrac{2}{\epsilon})\bigr].$$

Noticing moreover that

$$(\lambda - \beta)\nu\rho(r) + \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr]\bigr\} \\ = \nu\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr]\bigr\} + \nu\biggl[\int_{\beta}^{\lambda} \pi_{\exp(-\alpha r)}(r)d\alpha\biggr],$$

and choosing $\rho = \pi_{\exp(-\lambda r)}$, we have proved

Theorem 1.21 *For any positive real constants β, γ and η, such that $\gamma < \eta\bigl[\exp(\frac{\eta}{N}) - 1\bigr]^{-1}$, defining λ by condition (1.15), so that $\lambda = -N\log\Bigl\{1 - \frac{\gamma}{\eta}\bigl[\exp(\frac{\eta}{N}) - 1\bigr]\Bigr\}$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$, any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$\nu\biggl[\int_{\beta}^{\gamma} \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\bigl(\Phi_{-\frac{\eta}{N}}\circ R\bigr)d\alpha\biggr] \\ \leq \nu\biggl[\int_{\beta}^{\lambda} \pi_{\exp(-\alpha r)}(r)d\alpha\biggr] + \bigl(1+\tfrac{\beta}{\eta}\bigr)\bigl[\mathcal{K}(\nu,\overline{\mu}) + \log\bigl(\tfrac{2}{\epsilon}\bigr)\bigr].$$

Let us remark that this theorem does not require that $\beta < \gamma$, and thus provides both an upper and a lower bound for the quantity of interest:

Corollary 1.22 *For any positive real constants β, γ and η such that $\max\{\beta,\gamma\} < \eta\bigl[\exp(\frac{\eta}{N}) - 1\bigr]^{-1}$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\nu : \Omega \to \mathcal{M}_+^1(M)$ and $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$\nu\biggl[\int_{-N\log\{1-\frac{\beta}{N}[\exp(\frac{\eta}{N})-1]\}}^{\gamma} \pi_{\exp(-\alpha r)}(r)d\alpha\biggr] - \bigl(1+\tfrac{\gamma}{\eta}\bigr)\bigl[\mathcal{K}(\nu,\overline{\mu}) + \log\bigl(\tfrac{3}{\epsilon}\bigr)\bigr] \\ \leq \nu\biggl[\int_{\beta}^{\gamma} \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}}\circ R)}\bigl(\Phi_{-\frac{\eta}{N}}\circ R\bigr)d\alpha\biggr] \\ \leq \nu\biggl[\int_{\beta}^{-N\log\{1-\frac{\gamma}{\eta}[\exp(\frac{\eta}{N})-1]\}} \pi_{\exp(-\alpha r)}(r)d\alpha\biggr] \\ + \bigl(1+\tfrac{\beta}{\eta}\bigr)\bigl[\mathcal{K}(\nu,\overline{\mu}) + \log\bigl(\tfrac{3}{\epsilon}\bigr)\bigr].$$

We can then remember that

$$\mathcal{K}(\nu, \overline{\mu}) = \xi\bigl(\nu - \overline{\mu}\bigr)\left[\int_\beta^\gamma \pi_{\exp(-\alpha\Phi_{-\frac{\eta}{N}} \circ R)}\bigl(\Phi_{-\frac{\eta}{N}} \circ R\bigr) d\alpha\right] + \mathcal{K}(\nu, \mu) - \mathcal{K}(\overline{\mu}, \mu),$$

to conclude that, putting

$$G_\eta(\alpha) = -N \log\bigl\{1 - \frac{\alpha}{\eta}\bigl[\exp(\frac{\eta}{N}) - 1\bigr]\bigr\} \geq \alpha, \qquad \alpha \in \mathbb{R}_+, \tag{1.16}$$

and

$$\frac{d\widehat{\nu}}{d\mu}(m) \overset{\text{def}}{=} \frac{\exp\bigl[-h(m)\bigr]}{\mu\bigl[\exp(-h)\bigr]} \text{ where } h(m) = \xi \int_{G_\eta(\beta)}^\gamma \pi_{\exp(-\alpha r)}(m, r) d\alpha, \tag{1.17}$$

the divergence of ν with respect to the local prior $\overline{\mu}$ is bounded by

$$\begin{aligned}
&\bigl[1 - \xi\bigl(1 + \tfrac{\beta}{\eta}\bigr)\bigr]\mathcal{K}(\nu, \overline{\mu}) \\
&\quad \leq \xi\nu\left[\int_\beta^{G_\eta(\gamma)} \pi_{\exp(-\alpha r)}(r) d\alpha\right] - \xi\overline{\mu}\left[\int_{G_\eta(\beta)}^\gamma \pi_{\exp(-\alpha r)}(r) d\alpha\right] \\
&\qquad\qquad\qquad + \mathcal{K}(\nu, \mu) - \mathcal{K}(\overline{\mu}, \mu) + \xi\bigl(2 + \tfrac{\beta+\gamma}{\eta}\bigr)\log\bigl(\tfrac{3}{\epsilon}\bigr) \\
&\quad \leq \xi\nu\left[\int_\beta^{G_\eta(\gamma)} \pi_{\exp(-\alpha r)}(r) d\alpha\right] + \mathcal{K}(\nu, \mu) \\
&\qquad\qquad + \log\left\{\mu\left[\exp\left(-\xi\int_{G_\eta(\beta)}^\gamma \pi_{\exp(-\alpha r)}(r) d\alpha\right)\right]\right\} \\
&\qquad\qquad\qquad + \xi\bigl(2 + \tfrac{\beta+\gamma}{\eta}\bigr)\log\bigl(\tfrac{3}{\epsilon}\bigr) \\
&\quad = \mathcal{K}(\nu, \widehat{\nu}) + \xi\nu\left[\left(\int_\beta^{G_\eta(\beta)} + \int_\gamma^{G_\eta(\gamma)}\right)\pi_{\exp(-\alpha r)}(r) d\alpha\right] \\
&\qquad\qquad\qquad + \xi\bigl(2 + \tfrac{\beta+\gamma}{\eta}\bigr)\log\bigl(\tfrac{3}{\epsilon}\bigr).
\end{aligned}$$

We have proved

Theorem 1.23. *For any positive constants β, γ and η such that $\max\{\beta, \gamma\} < \eta\bigl[\exp(\frac{\eta}{N}) - 1\bigr]^{-1}$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$ and any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$\mathcal{K}(\nu, \overline{\mu}) \leq \left[1 - \xi\left(1 + \frac{\beta}{\eta}\right)\right]^{-1}\Bigl\{\mathcal{K}(\nu, \widehat{\nu})$$

$$+ \xi \nu \biggl[\biggl(\int_{\beta}^{G_{\eta}(\beta)} + \int_{\gamma}^{G_{\eta}(\gamma)} \biggr) \pi_{\exp(-\alpha r)}(r) d\alpha \biggr] \\ + \xi \bigl(2 + \tfrac{\beta + \gamma}{\eta} \bigr) \log \bigl(\tfrac{3}{\epsilon} \bigr) \biggr\}$$

$$\leq \biggl[1 - \xi \Bigl(1 + \frac{\beta}{\eta} \Bigr) \biggr]^{-1} \biggl\{ \mathcal{K}(\nu, \widehat{\nu}) \\ + \xi \nu \biggl[\bigl[G_{\eta}(\gamma) - \gamma + G_{\eta}(\beta) - \beta \bigr] r^{\star} + \log \biggl(\frac{G_{\eta}(\beta) G_{\eta}(\gamma)}{\beta \gamma} \biggr) d_e \biggr] \\ + \xi \bigl(2 + \tfrac{\beta + \gamma}{\eta} \bigr) \log \bigl(\tfrac{3}{\epsilon} \bigr) \biggr\},$$

where the local prior $\overline{\mu}$ is defined by equation (1.13) *on page 35 and the local posterior $\widehat{\nu}$ and the function G_{η} are defined by equation* (1.17) *above.*

We can then use this theorem to give a local version of Theorem 1.20 (page 34). To get something pleasing to read, we can apply Theorem 1.23 with constants β', γ' and η chosen so that $\frac{2\xi}{1 - \xi(1 + \frac{\beta'}{\eta})} = 1$, $G_{\eta}(\beta') = \beta$ and $\gamma' = \lambda$, where β and λ are the constants appearing in Theorem 1.20. This gives

THEOREM 1.24. *For any positive real constants $\beta < \lambda$ and η such that $\lambda < \eta \bigl[\exp(\frac{\eta}{N}) - 1 \bigr]^{-1}$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distribution $\nu : \Omega \rightarrow \mathcal{M}_{+}^{1}(M)$, for any conditional posterior distribution $\rho : \Omega \times M \rightarrow \mathcal{M}_{+}^{1}(\Theta)$,*

$$\lambda \Phi_{\frac{\lambda}{N}} \bigl[\nu \rho(R) \bigr] - \beta \Phi_{-\frac{\beta}{N}} \bigl[\nu \rho(R) \bigr] \leq B_3(\nu, \rho), \ \textit{where}$$

$$B_3(\nu, \rho) = \nu \biggl[\int_{G_{\eta}^{-1}(\beta)}^{G_{\eta}(\lambda)} \pi_{\exp(-\alpha r)}(r) d\alpha \biggr] \\ + \Bigl(3 + \tfrac{G_{\eta}^{-1}(\beta)}{\eta} \Bigr) \mathcal{K} \bigl[\nu, \mu_{\exp \bigl[- \bigl(3 + \frac{G_{\eta}^{-1}(\beta)}{\eta} \bigr)^{-1} \int_{\beta}^{\lambda} \pi_{\exp(-\alpha r)}(r) d\alpha \bigr]} \bigr] \\ + \nu \bigl\{ \mathcal{K}(\rho, \pi_{\exp(-\lambda r)} \bigr] \bigr\} + \Bigl(4 + \tfrac{G_{\eta}^{-1}(\beta) + \lambda}{\eta} \Bigr) \log \bigl(\tfrac{4}{\epsilon} \bigr)$$

$$\leq \nu \biggl[\bigl[G_{\eta}(\lambda) - G_{\eta}^{-1}(\beta) \bigr] r^{\star} + \log \Bigl(\tfrac{G_{\eta}(\lambda)}{G_{\eta}^{-1}(\beta)} \Bigr) d_e \biggr] \\ + \Bigl(3 + \tfrac{G_{\eta}^{-1}(\beta)}{\eta} \Bigr) \mathcal{K} \bigl[\nu, \mu_{\exp \bigl[- \bigl(3 + \frac{G_{\eta}^{-1}(\beta)}{\eta} \bigr)^{-1} \int_{\beta}^{\lambda} \pi_{\exp(-\alpha r)}(r) d\alpha \bigr]} \bigr] \\ + \nu \bigl\{ \mathcal{K}(\rho, \pi_{\exp(-\lambda r)} \bigr] \bigr\} + \Bigl(4 + \tfrac{G_{\eta}^{-1}(\beta) + \lambda}{\eta} \Bigr) \log \bigl(\tfrac{4}{\epsilon} \bigr),$$

and where the function G_{η} is defined by equation (1.16) *on page 38.*

A first remark: if we had the stamina to use Cauchy Schwarz inequalities (or more generally Hölder inequalities) on exponential moments instead of using weighted union bounds on deviation inequalities, we could have replaced $\log(\frac{4}{\epsilon})$ with $-\log(\epsilon)$ in the above inequalities.

We see that we have achieved the desired kind of localization of Theorem 1.20 (page 34), since the new empirical entropy term

$$\mathcal{K}[\nu, \mu_{\exp[-\xi \int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha]}]$$

cancels for a value of the posterior distribution on the index set ν which is of the same form as the one minimizing the bound $B_1(\nu, \rho)$ of Theorem 1.20 (with a decreased constant, as could be expected). In a typical parametric setting, we will have

$$\int_\beta^\lambda \pi_{\exp(-\alpha r)}(r) d\alpha \simeq (\lambda - \beta) r^\star(m) + \log\left(\tfrac{\lambda}{\beta}\right) d_e(m),$$

and therefore, if we choose for ν the Dirac mass at

$$\widehat{m} \in \arg\min_{m \in M} r^\star(m) + \tfrac{\log(\frac{\lambda}{\beta})}{\lambda - \beta} d_e(m),$$

and $\rho(m, \cdot) = \pi_{\exp(-\lambda r)}(m, \cdot)$, we will get, in the case when the index set M is countable,

$$\begin{aligned} B_3(\nu, \rho) \lesssim \max \Biggl\{ \bigl[G_\eta(\lambda) - G_\eta^{-1}(\beta)\bigr], (\lambda - \beta) \frac{\log\Bigl[\frac{G_\eta(\lambda)}{G_\eta^{-1}(\beta)}\Bigr]}{\log(\frac{\lambda}{\beta})} \Biggr\} \\ \times \Bigl[r^\star(\widehat{m}) + \tfrac{\log(\frac{\lambda}{\beta})}{\lambda - \beta} d_e(\widehat{m}) \Bigr] \\ + \Bigl(3 + \tfrac{G_\eta^{-1}(\beta)}{\eta}\Bigr) \log \Biggl\{ \sum_{m \in M} \tfrac{\mu(m)}{\mu(\widehat{m})} \exp\Biggl[- \Bigl(3 + \tfrac{G_\eta^{-1}(\beta)}{\eta}\Bigr)^{-1} \\ \times \Bigl\{ (\lambda - \beta)\bigl[r^\star(m) - r^\star(\widehat{m})\bigr] + \log\bigl(\tfrac{\lambda}{\beta}\bigr)\bigl[d_e(m) - d_e(\widehat{m})\bigr] \Bigr\} \Biggr] \Biggr\} \\ + \Bigl(4 + \tfrac{G_\eta^{-1}(\beta) + \lambda}{\eta}\Bigr) \log\bigl(\tfrac{4}{\epsilon}\bigr). \end{aligned}$$

Therefore, as long as there are not too many of them, we do not feel strongly in this bound the models for which the penalized minimum empirical risk $r^\star(m) + \frac{\log(\frac{\lambda}{\beta})}{\lambda - \beta} d_e(m)$ is far from optimal.

1.4. Relative bounds. The behaviour of the minimum of the empirical process $\theta \mapsto r(\theta)$ is known to depend on the covariances between pairs

$\bigl[r(\theta), r(\theta')\bigr]$, $\theta, \theta' \in \Theta$. Accordingly, our previous study, based on the analysis of the variance of $r(\theta)$ (or technically on some exponential moment playing quite the same role), is missing some accuracy in some circumstances (namely when $\inf_\Theta R$ is not close enough to zero). In this subsection, instead of bounding the expected risk $\rho(R)$, we are going to upper bound the difference $\rho(R) - \inf_\Theta R$, and more generally $\rho(R) - R(\widetilde{\theta})$, where $\widetilde{\theta} \in \Theta$ is some fixed parameter value. Eventually in the next subsection we will analyze $\rho(R) - \pi_{\exp(-\beta R)}(R)$, allowing to compare the expected error rate of a posterior distribution ρ with the error rate of a Gibbs prior distribution. Thus relative bounds are not exactly of the same nature as previous ones: although it is not possible to estimate $\rho(R)$ with an order of precision higher than $(\rho(R)/N)^{1/2}$, it is still possible in some situations to reach a better precision for $\rho(R) - \inf_\Theta R$, as we will see. The study of PAC-Bayesian relative bounds stems from the second and third part of J. Y. Audibert's dissertation [3].

We will suggest two different kinds of applications of these bounds. The first more obvious one is to upper bound $\rho(R) - \inf_\Theta R$ to get an idea of the performance of the posterior distribution ρ.

The second application is to compare the classification model indexed by Θ with a submodel indexed by one of its measurable subsets $\Theta_1 \subset \Theta$. For this purpose we are going to compare $\rho(R)$, where $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ is any posterior distribution, with $R(\widetilde{\theta})$, where $\widetilde{\theta} \in \Theta_1$ is some possibly unobservable value of the parameter in the submodel defined by Θ_1. We will typically consider the case when $\widetilde{\theta} \in \arg\min_{\Theta_1} R$. In this special case, a negative bound for $\rho(R) - R(\widetilde{\theta}) = \rho(R) - \inf_{\Theta_1} R$ indicates that it is definitely worth using a randomized estimator ρ supported by the larger parameter set Θ instead of using only the classification model defined by the smaller set Θ_1.

1.4.1. Basic inequalities. Relative bounds in this section are based on the control of $r(\theta) - r(\widetilde{\theta})$, where $\theta, \widetilde{\theta} \in \Theta$. These differences are related to the random variables

$$\psi_i(\theta, \widetilde{\theta}) = \sigma_i(\theta) - \sigma_i(\widetilde{\theta}) = \mathbb{1}\bigl[f_\theta(X_i) \neq Y_i\bigr] - \mathbb{1}\bigl[f_{\widetilde{\theta}}(X_i) \neq Y_i\bigr].$$

Some supplementary technical difficulties, as compared to the previous sections, come from the fact that $\psi_i(\theta, \widetilde{\theta})$ takes three values, whereas $\sigma_i(\theta)$ takes only two. Let $r'(\theta, \widetilde{\theta}) = r(\theta) - r(\widetilde{\theta})$ and $R'(\theta, \widetilde{\theta}) = R(\theta) - R(\widetilde{\theta})$. We have as usual from independence that

$$\log\Bigl\{\mathbb{P}\Bigl[\exp\bigl[-\lambda r'(\theta,\widetilde{\theta})\bigr]\Bigr]\Bigr\} = \sum_{i=1}^N \log\Bigl\{\mathbb{P}\Bigl[\exp\bigl[-\tfrac{\lambda}{N}\psi_i(\theta,\widetilde{\theta})\bigr]\Bigr]\Bigr\}$$
$$\leq N\log\biggl\{\frac{1}{N}\sum_{i=1}^N \mathbb{P}\Bigl\{\exp\Bigl[-\frac{\lambda}{N}\psi_i(\theta,\widetilde{\theta})\Bigr]\Bigr\}\biggr\}.$$

Let C_i be the distribution of $\psi_i(\theta,\widetilde{\theta})$ under $\mathbb{P}$ and let $\bar{C} = \frac{1}{N}\sum_{i=1}^N C_i \in \mathcal{M}_+^1\bigl(\{-1,0,1\}\bigr)$. With these notations

$$\log\Bigl\{\mathbb{P}\Bigl[\exp\bigl[-\lambda r'(\theta,\widetilde{\theta})\bigr]\Bigr]\Bigr\} \leq N\log\biggl\{\int \exp\Bigl(-\frac{\lambda}{N}\psi\Bigr)\bar{C}(d\psi)\biggr\}. \tag{1.18}$$

The right-hand side of this inequality is a function of $\bar{C}$. On the other hand, $\bar{C}$ being a probability measure on a three point set, is defined by two parameters, that we may take equal to $\int \psi\bar{C}(d\psi)$ and $\int \psi^2\bar{C}(d\psi)$. To this purpose, let us introduce

$$M'(\theta,\widetilde{\theta}) = \int \psi^2\bar{C}(d\psi) = \bar{C}(+1)+\bar{C}(-1) = \frac{1}{N}\sum_{i=1}^N \mathbb{P}\bigl[\psi_i^2(\theta,\widetilde{\theta})\bigr], \quad \theta,\widetilde{\theta}\in\Theta.$$

It is a pseudo distance (meaning that it is symmetric and satisfies the triangle inequality), since it can also be written as

$$M'(\theta,\widetilde{\theta}) = \frac{1}{N}\sum_{i=1}^N \mathbb{P}\Bigl\{\Bigl|\mathbb{1}\bigl[f_\theta(X_i)\neq Y_i\bigr] - \mathbb{1}\bigl[f_{\widetilde{\theta}}(X_i)\neq Y_i\bigr]\Bigr|\Bigr\}, \quad \theta,\widetilde{\theta}\in\Theta.$$

It is readily seen that

$$N\log\left\{\int \exp\left(-\frac{\lambda}{N}\psi\right)\bar{C}(d\psi)\right\} = -\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\theta,\widetilde{\theta}),M'(\theta,\widetilde{\theta})\bigr],$$

where

$$\Psi_a(p,m) = -a^{-1}\log\Bigl[(1-m)+\frac{m+p}{2}\exp(-a)+\frac{m-p}{2}\exp(a)\Bigr]$$
$$= -a^{-1}\log\Bigl\{1-\sinh(a)\bigl[p-m\tanh(\tfrac{a}{2})\bigr]\Bigr\}.$$

Thus plugging this equality into inequality (1.18) we see that for any real parameter λ,

$$\log\Bigl\{\mathbb{P}\Bigl[\exp\bigl[-\lambda r'(\theta,\widetilde{\theta})\bigr]\Bigr]\Bigr\} \leq -\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\theta,\widetilde{\theta}),M'(\theta,\widetilde{\theta})\bigr],$$

To make a link with previous works initiated by Mammen and Tsybakov (see e.g. [28, 34]), we may consider the pseudo distance D on Θ defined on page 14 by equation (1.2). This distance only depends on the distribution of the patterns. It is often used to formulate margin assumptions (in the sense of Mammen and Tsybakov). Here we are going to work rather with M': as it is dominated by D in the sense that $M'(\theta, \widetilde{\theta}) \leq D(\theta, \widetilde{\theta})$, $\theta, \widetilde{\theta} \in \Theta$, with equality in the important case of binary classification, hypotheses formulated on D induce hypotheses on M', and working with M' may only sharpen the results when compared to working with D.

Using the same reasoning as in the previous section, we deduce

Theorem 1.25. *For any real parameter λ, any $\widetilde{\theta} \in \Theta$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \Bigl[\rho\bigl\{\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot, \widetilde{\theta}\,), M'(\cdot, \widetilde{\theta}\,)\bigr]\bigr\} - \rho\bigl[r'(\cdot, \widetilde{\theta})\bigr]\Bigr] - \mathcal{K}(\rho, \pi)\biggr]\biggr\} \leq 1.$$

We are now going to derive some variant of Theorem 1.25. In this theorem, we obtain an inequality comparing one observed quantity $\rho\bigl[r'(\cdot, \widetilde{\theta}\,)\bigr]$ with two unobversed ones, $\rho\bigl[R'(\cdot, \widetilde{\theta}\,)\bigr]$ and $\rho\bigl[M'(\cdot, \widetilde{\theta}\,)\bigr]$ (because of the convexity of the function $\lambda\Psi_{\frac{\lambda}{N}}$,

$$\lambda\rho\bigl\{\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot, \widetilde{\theta}\,), M'(\cdot, \widetilde{\theta}\,)\bigr]\bigr\} \geq \lambda\Psi_{\frac{\lambda}{N}}\bigl\{\rho\bigl[R'(\cdot, \widetilde{\theta}\,)\bigr], \rho\bigl[M'(\cdot, \widetilde{\theta}\,)\bigr]\bigr\}.)$$

This may be inconvenient when looking for an empirical bound for $\rho\bigl[R'(\cdot, \widetilde{\theta})\bigr]$, and we are going now to seek an inequality comparing $\rho\bigl[R'(\cdot, \widetilde{\theta}\,)\bigr]$ with empirical quantities only. This is possible through a change of variables in the exponential inequality. Indeed, if we consider now random variables $\chi_i(\theta, \widetilde{\theta})$, such that

$$1 - \frac{\lambda}{N}\psi_i = \exp\left(-\frac{\lambda}{N}\chi_i\right),$$

which is possible when $\frac{\lambda}{N} \in \,)-1, 1($ and leads to define

$$\chi_i = -\frac{N}{\lambda}\log\left(1 - \frac{\lambda}{N}\psi_i\right),$$

we obtain easily following the same reasoning as previously

$$\log\biggl\{\mathbb{P}\biggl\{\exp\biggl[\sum_{i=1}^N \log\Bigl(1 - \frac{\lambda}{N}\psi_i\Bigr)\biggr]\biggr\}\biggr\}$$

$$\leq \sum_{i=1}^N \log \Bigl[1 - \frac{\lambda}{N} \mathbb{P}(\psi_i)\Bigr] \leq N \log \Bigl[1 - \frac{\lambda}{N} R'(\theta, \widetilde{\theta}\,)\Bigr].$$

Let us replace for simplicity λ/N with λ. Let us also introduce the random pseudo distance

$$\begin{aligned} m'(\theta, \widetilde{\theta}) &= \frac{1}{N} \sum_{i=1}^N \psi_i(\theta, \widetilde{\theta})^2 \\ &= \frac{1}{N} \sum_{i=1}^N \Bigl| \mathbb{1}\bigl[f_\theta(X_i) \neq Y_i\bigr] - \mathbb{1}\bigl[f_{\widetilde{\theta}}(X_i) \neq Y_i\bigr] \Bigr|, \quad \theta, \widetilde{\theta} \in \Theta. \quad (1.19) \end{aligned}$$

This is the empirical counter part of M', since $\mathbb{P}(m') = M'$. Let us notice that

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \log\bigl[1 - \lambda \psi_i(\theta, \widetilde{\theta})\bigr] = \frac{\log(1-\lambda) - \log(1+\lambda)}{2} r'(\theta, \widetilde{\theta}) \\ + \frac{\log(1-\lambda) + \log(1+\lambda)}{2} m'(\theta, \widetilde{\theta}) \\ = \frac{1}{2} \log\left(\frac{1-\lambda}{1+\lambda}\right) r'\bigl(\theta, \widetilde{\theta}\,\bigr) + \frac{1}{2} \log(1-\lambda^2) m'\bigl(\theta, \widetilde{\theta}\,\bigr). \end{aligned}$$

With these notations, we can conveniently write the previous inequality as

$$\mathbb{P}\Biggl\{ \exp\Biggl[-N \log\bigl[1 - \lambda R'(\theta, \widetilde{\theta})\bigr] \\ - \frac{N}{2} \log\left(\frac{1+\lambda}{1-\lambda}\right) r'\bigl(\theta, \widetilde{\theta}\,\bigr) + \frac{N}{2} \log\bigl(1 - \lambda^2\bigr) m'\bigl(\theta, \widetilde{\theta}\,\bigr) \Biggr] \Biggr\} \leq 1.$$

Integrating with respect to a prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$, we obtain

Theorem 1.26. *For any real parameter $\lambda \in)-1, 1($, for any $\widetilde{\theta} \in \Theta$, for any prior probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\Biggl\{ \exp\Biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \biggl\{ -N\rho\Bigl\{ \log\bigl[1 - \lambda R'(\cdot, \widetilde{\theta}\,)\bigr] \Bigr\}$$

$$- \frac{N}{2}\log\left(\frac{1+\lambda}{1-\lambda}\right)\rho\bigl[r'(\cdot,\widetilde{\theta}\,)\bigr]$$
$$+ \frac{N}{2}\log(1-\lambda^2)\rho\bigl[m'(\cdot,\widetilde{\theta}\,)\bigr] - \mathcal{K}(\rho,\pi)\Bigr\}\Bigr]\Bigr\} \leq 1.$$

1.4.2. Non random bounds. Let us first deduce a non random bound from Theorem 1.25. This theorem can be conveniently taken advantage of by throwing the non linearity into a localized prior, considering the prior probability measure μ defined by

$$\frac{d\mu}{d\pi}(\theta) = \frac{\exp\bigl\{-\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\theta,\widetilde{\theta}\,),M'(\theta,\widetilde{\theta}\,)\bigr] + \beta R'(\theta,\widetilde{\theta}\,)\bigr\}}{\pi\Bigl\{\exp\bigl\{-\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot,\widetilde{\theta}\,),M'(\cdot,\widetilde{\theta}\,)\bigr] + \beta R'(\cdot,\widetilde{\theta}\,)\bigr\}\Bigr\}}.$$

Indeed, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\mathcal{K}(\rho,\mu) = \mathcal{K}(\rho,\pi) + \lambda\rho\Bigl\{\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot,\widetilde{\theta}\,),M'(\cdot,\widetilde{\theta}\,)\bigr]\Bigr\} - \beta\rho\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]$$
$$+ \log\Bigl\{\pi\Bigl[\exp\bigl\{-\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot,\widetilde{\theta}\,),M'(\cdot,\widetilde{\theta}\,)\bigr] + \beta R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\}\Bigr]\Bigr\}.$$

Plugging this into Theorem 1.25 and using the convexity of the exponential function, we see that for any posterior probability distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\beta\mathbb{P}\bigl\{\rho\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\} \leq \lambda\mathbb{P}\bigl\{\rho\bigl[r'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\} + \mathbb{P}\bigl[\mathcal{K}(\rho,\pi)\bigr]$$
$$+ \log\Bigl\{\pi\Bigl[\exp\bigl\{-\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot,\widetilde{\theta}\,),M'(\cdot,\widetilde{\theta}\,)\bigr] + \beta R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\}\Bigr]\Bigr\}.$$

We can then recall that

$$\lambda\rho\bigl[r'(\cdot,\widetilde{\theta}\,)\bigr] + \mathcal{K}(\rho,\pi) = \mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda r)}\bigr] - \log\Bigl\{\pi\Bigl[\exp\bigl[-\lambda r'(\cdot,\widetilde{\theta}\,)\bigr]\Bigr]\Bigr\},$$

and notice moreover that

$$-\mathbb{P}\Bigl\{\log\Bigl\{\pi\Bigl[\exp\bigl[-\lambda r'(\cdot,\widetilde{\theta}\,)\bigr]\Bigr]\Bigr\}\Bigr\} \leq -\log\Bigl\{\pi\Bigl[\exp\bigl[-\lambda R'(\cdot,\widetilde{\theta}\,)\bigr]\Bigr]\Bigr\},$$

since $R' = \mathbb{P}(r')$ and $h \mapsto \log\Bigl\{\pi\bigl[\exp(h)\bigr]\Bigr\}$ is a convex functional. Putting these two remarks together, we obtain

THEOREM 1.27. *For any real positive parameter λ, for any prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}\mathbb{P}\bigl\{\rho\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\} &\leq \frac{1}{\beta}\mathbb{P}\bigl[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\bigr] \\ &\quad + \frac{1}{\beta}\log\Bigl\{\pi\Bigl[\exp\bigl\{-\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot,\widetilde{\theta}\,),M'(\cdot,\widetilde{\theta}\,)\bigr]+\beta R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\}\Bigr]\Bigr\} \\ &\qquad - \frac{1}{\beta}\log\Bigl\{\pi\Bigl[\exp\bigl[-\lambda R'(\cdot,\widetilde{\theta}\,)\bigr]\Bigr]\Bigr\} \\ &\leq \frac{1}{\beta}\mathbb{P}\bigl[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\bigr] \\ &\quad + \frac{1}{\beta}\log\Bigl\{\pi\Bigl[\exp\bigl\{-\bigl[N\sinh(\tfrac{\lambda}{N})-\beta\bigr]R'(\cdot,\widetilde{\theta}\,) \\ &\qquad + 2N\sinh(\tfrac{\lambda}{2N})^2 M'(\cdot,\widetilde{\theta}\,)\bigr\}\Bigr]\Bigr\} \\ &\qquad - \frac{1}{\beta}\log\Bigl\{\pi\Bigl[\exp\bigl[-\lambda R'(\cdot,\widetilde{\theta}\,)\bigr]\Bigr]\Bigr\}.\end{aligned}$$

It may be interesting to derive some more suggestive (but slightly weaker) bound in the important case when $\Theta_1 = \Theta$ and $R(\widetilde{\theta}) = \inf_\Theta R$. In this case, it is convenient to introduce the *margin function*

$$\varphi(x) = \sup_{\theta\in\Theta} M'(\theta,\widetilde{\theta}) - xR'(\theta,\widetilde{\theta}), \quad x \in \mathbb{R}_+. \tag{1.20}$$

We see that φ is convex and nonnegative on $\mathbb{R}_+$. Using the bound $M'(\theta,\widetilde{\theta}\,) \leq xR'(\theta,\widetilde{\theta}\,) + \varphi(x)$, we obtain

$$\begin{aligned}\mathbb{P}\bigl\{\rho\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\} &\leq \frac{1}{\beta}\mathbb{P}\bigl[\mathcal{K}(\rho,\pi_{\exp(-\lambda r)})\bigr] \\ &+ \frac{1}{\beta}\log\Bigl\{\pi\Bigl[\exp\Bigl\{-\bigl\{N\sinh(\tfrac{\lambda}{N})\bigl[1-x\tanh(\tfrac{\lambda}{2N})\bigr]-\beta\bigr\}R'(\cdot,\widetilde{\theta}\,)\Bigr\}\Bigr]\Bigr\} \\ &\qquad + \frac{N\sinh(\frac{\lambda}{N})\tanh(\frac{\lambda}{2N})}{\beta}\varphi(x) - \frac{1}{\beta}\log\Bigl\{\pi\Bigl[\exp\bigl[-\lambda R'(\cdot,\widetilde{\theta}\,)\bigr]\Bigr]\Bigr\}.\end{aligned}$$

Let us make the change of variable $\gamma = N\sinh(\frac{\lambda}{N})\bigl[1-x\tanh(\frac{\lambda}{2N})\bigr]-\beta$ to obtain

COROLLARY 1.28. *For any real positive parameters x, γ and λ such that $x \leq \tanh(\frac{\lambda}{2N})^{-1}$ and $0 \leq \gamma < N\sinh(\frac{\lambda}{N})\bigl[1-x\tanh(\frac{\lambda}{2N})\bigr]$,*

$$\mathbb{P}\bigl[\rho(R)\bigr] - \inf_{\Theta} R \leq \Bigl\{N \sinh(\tfrac{\lambda}{N})\bigl[1 - x \tanh(\tfrac{\lambda}{2N})\bigr] - \gamma\Bigr\}^{-1}$$

$$\times \Biggl\{\int_{\gamma}^{\lambda} \bigl[\pi_{\exp(-\alpha R)}(R) - \inf_{\Theta} R\bigr] d\alpha$$

$$+ N \sinh\bigl(\tfrac{\lambda}{N}\bigr) \tanh\bigl(\tfrac{\lambda}{2N}\bigr)\varphi(x) + \mathbb{P}\bigl[\mathcal{K}(\rho, \pi_{\exp(-\lambda r)})\bigr]\Biggr\}.$$

Let us remark that these results, although well suited to study Mammen and Tsybakov's margin assumptions, hold in the general case: introducing the convex *expected margin function* φ is a substitute for making hypotheses about the relations between R and D.

Using the fact that $R'(\theta, \widetilde{\theta}) \geq 0$, $\theta \in \Theta$ and that $\varphi(x) \geq 0$, $x \in \mathbb{R}_+$, we can weaken and simplify even more the preceding corollary to get

COROLLARY 1.29. *For any real parameters β, λ and x such that $x \geq 0$ and $0 \leq \beta < \lambda - x\frac{\lambda^2}{2N}$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\bigl[\rho(R)\bigr] \leq \inf_{\Theta} R$$

$$+ \Bigl[\lambda - x\tfrac{\lambda^2}{2N} - \beta\Bigr]^{-1}\Biggl\{\int_{\beta}^{\lambda} \bigl[\pi_{\exp(-\alpha R)}(R) - \inf_{\Theta} R\bigr] d\alpha$$

$$+ \mathbb{P}\bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr]\bigr\} + \varphi(x)\frac{\lambda^2}{2N}\Biggr\}.$$

Let us apply this bound under the *margin assumption* first considered by Mammen and Tsybakov [28, 34], which tells that for some real positive constant c and some real exponent $\kappa \geq 1$,

$$R'(\theta, \widetilde{\theta}) \geq c D(\theta, \widetilde{\theta})^{\kappa}, \qquad \theta \in \Theta. \tag{1.21}$$

In the case when $\kappa = 1$, then $\varphi(c^{-1}) = 0$, proving that

$$\mathbb{P}\bigl\{\pi_{\exp(-\lambda r)}\bigl[R'(\cdot, \widetilde{\theta})\bigr]\bigr\} \leq \frac{\int_{\beta}^{\lambda} \pi_{\exp(-\gamma R)}\bigl[R'(\cdot, \widetilde{\theta})\bigr] d\gamma}{N \sinh(\frac{\lambda}{N})\bigl[1 - c^{-1} \tanh(\frac{\lambda}{2N})\bigr] - \beta}$$

$$\leq \frac{\int_{\beta}^{\lambda} \pi_{\exp(-\gamma R)}\bigl[R'(\cdot, \widetilde{\theta})\bigr] d\gamma}{\lambda - \frac{\lambda^2}{2cN} - \beta}.$$

Taking for example $\lambda = \frac{cN}{2}$, $\beta = \frac{\lambda}{2} = \frac{cN}{4}$, we obtain

$$\mathbb{P}\bigl[\pi_{\exp(-2^{-1}cNr)}(R)\bigr] \leq \inf R + \frac{8}{cN}\int_{\frac{cN}{4}}^{\frac{cN}{2}} \pi_{\exp(-\gamma R)}\bigl[R'(\cdot, \widetilde{\theta})\bigr] d\gamma$$

$$\leq \inf R + 2\pi_{\exp(-\frac{cN}{4}R)}\bigl[R'(\cdot, \widetilde{\theta})\bigr].$$

If moreover the behaviour of the prior distribution π is parametric meaning that $\pi_{\exp(-\beta R)}\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr] \leq \frac{d}{\beta}$, for some positive real constant d linked with the dimension of the classification model, then

$$\mathbb{P}\bigl[\pi_{\exp(-\frac{cN}{2}r)}(R)\bigr] \leq \inf R + \frac{8\log(2)d}{cN} \leq \inf R + \frac{5.55\,d}{cN}.$$

In the case when $\kappa > 1$,

$$\varphi(x) \leq (\kappa-1)\kappa^{-\frac{\kappa}{\kappa-1}}(cx)^{-\frac{1}{\kappa-1}} = (1-\kappa^{-1})(\kappa c x)^{-\frac{1}{\kappa-1}},$$

thus $\mathbb{P}\bigl\{\pi_{\exp(-\lambda r)}\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\}$

$$\leq \frac{\int_\beta^\lambda \pi_{\exp(-\gamma R)}\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]d\gamma + (1-\kappa^{-1})(\kappa c x)^{-\frac{1}{\kappa-1}}\frac{\lambda^2}{2N}}{\lambda - \frac{x\lambda^2}{2N} - \beta}.$$

Taking for instance $\beta = \frac{\lambda}{2}$, $x = \frac{N}{2\lambda}$, and putting $b = (1-\kappa^{-1})(c\kappa)^{-\frac{1}{\kappa-1}}$, we obtain

$$\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr] - \inf R \leq \frac{4}{\lambda}\int_{\lambda/2}^{\lambda}\pi_{\exp(-\gamma R)}\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]d\gamma + b\left(\frac{2\lambda}{N}\right)^{\frac{\kappa}{\kappa-1}}.$$

In the *parametric* case when $\pi_{\exp(-\gamma R)}\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr] \leq \frac{d}{\gamma}$, we get

$$\mathbb{P}\bigl[\pi_{\exp(-\lambda r)}(R)\bigr] - \inf R \leq \frac{4\log(2)d}{\lambda} + b\left(\frac{2\lambda}{N}\right)^{\frac{\kappa}{\kappa-1}}.$$

Taking

$$\overline{\lambda} = 2^{-1}\bigl[8\log(2)d\bigr]^{\frac{\kappa-1}{2\kappa-1}}(\kappa c)^{\frac{1}{2\kappa-1}}N^{\frac{\kappa}{2\kappa-1}},$$

we obtain

$$\mathbb{P}\bigl[\pi_{\exp(-\overline{\lambda} r)}(R)\bigr] - \inf R \leq (2-\kappa^{-1})(\kappa c)^{-\frac{1}{2\kappa-1}}\left(\frac{8\log(2)d}{N}\right)^{\frac{\kappa}{2\kappa-1}}.$$

We see that this formula coincides with the result for $\kappa = 1$. We can thus reduce the two cases to a single one and state

Corollary 1.30. *Let us assume that for some $\widetilde{\theta} \in \Theta$, some positive real constant c, some real exponent $\kappa \geq 1$ and for any $\theta \in \Theta$, $R(\theta) \geq R(\widetilde{\theta}) + cD(\theta,\widetilde{\theta})^\kappa$. Let us also assume that for some positive real constant d and any positive real parameter γ, $\pi_{\exp(-\gamma R)}(R) - \inf R \leq \frac{d}{\gamma}$. Then*

$$\mathbb{P}\Bigl[\pi_{\exp\bigl\{-2^{-1}[8\log(2)d]^{\frac{\kappa-1}{2\kappa-1}}(\kappa c)^{\frac{1}{2\kappa-1}}N^{\frac{\kappa}{2\kappa-1}}r\bigr\}}(R)\Bigr] \\ \leq \inf R + (2-\kappa^{-1})(\kappa c)^{-\frac{1}{2\kappa-1}}\left(\frac{8\log(2)d}{N}\right)^{\frac{\kappa}{2\kappa-1}}.$$

Let us remark that the exponent of N is this corollary is known to be the minimax exponent under these assumptions: it is unimprovable, whatever estimator is used in place of the Gibbs posterior shown here (at least in the worst case compatible with the hypotheses). The interest of the corollary is to show not only the minimax exponent in N, but also an explicit non asymptotic bound with reasonable and simple constants. It is also clear that we could have got slightly better constants if we had kept the full strength of Theorem 1.27 (page 46) instead of using the weaker Corollary 1.29 (page 47).

We will prove in the following empirical bounds showing how the constant λ can be estimated from the data instead of being chosen according to some margin and complexity assumptions.

1.4.3. Unbiased empirical bounds. We are going to provide an empirical counter part for the *expected margin function* φ. It will appear in empirical bounds having otherwise the same structure as the non random bound we just proved. Anyhow, we will not launch into trying to compare the behaviour of our proposed *empirical margin function* with the *expected margin function*, since the margin function involves taking a supremum which is not straightforward to handle.

Let us start as in the previous subsection with the inequality

$$\beta\mathbb{P}\Bigl\{\rho\bigl[R'(\cdot,\widetilde{\theta})\bigr]\Bigr\} \leq \mathbb{P}\Bigl\{\lambda\rho\bigl[r'(\cdot,\widetilde{\theta})\bigr] + \mathcal{K}(\rho,\pi)\Bigr\} \\ + \log\Bigl\{\pi\Bigl[\exp\bigl\{-\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot,\widetilde{\theta}), M'(\cdot,\widetilde{\theta})\bigr] + \beta R'(\cdot,\widetilde{\theta})\bigr\}\Bigr]\Bigr\}.$$

We have already defined by equation (1.19) the empirical pseudo distance

$$m'(\theta,\widetilde{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\psi_i(\theta,\widetilde{\theta})^2.$$

Recalling that $\mathbb{P}\bigl[m'(\theta,\widetilde{\theta})\bigr] = M'(\theta,\widetilde{\theta})$, and using the convexity of $h \mapsto \log\Bigl\{\pi\bigl[\exp(h)\bigr]\Bigr\}$, leads to the following inequalities:

$$\begin{aligned}\log\Bigl\{\pi\Bigl[\exp\bigl\{-\lambda\Psi_{\frac{\lambda}{N}}\bigl[R'(\cdot,\widetilde{\theta}\,),M'(\cdot,\widetilde{\theta}\,)\bigr]+\beta R'(\cdot,\widetilde{\theta}\,)\bigr\}\Bigr]\Bigr\}\\ \leq \log\Bigl\{\pi\Bigl[\exp\bigl\{-N\sinh(\tfrac{\lambda}{N})R'(\cdot,\widetilde{\theta}\,)\\ + N\sinh(\tfrac{\lambda}{N})\tanh(\tfrac{\lambda}{2N})M'(\cdot,\widetilde{\theta}\,)+\beta R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\}\Bigr]\Bigr\}\\ \leq \mathbb{P}\biggl\{\log\Bigl\{\pi\Bigl[\exp\bigl\{-\bigl[N\sinh(\tfrac{\lambda}{N})-\beta\bigr]r'(\cdot,\widetilde{\theta}\,)\\ + N\sinh(\tfrac{\lambda}{N})\tanh(\tfrac{\lambda}{2N})m'(\cdot,\widetilde{\theta}\,)\bigr\}\Bigr]\Bigr\}\biggr\}.\end{aligned}$$

We may moreover remark that

$$\begin{aligned}\lambda\rho\bigl[r'(\cdot,\widetilde{\theta}\,)\bigr]+\mathcal{K}(\rho,\pi)=\bigl[\beta-N\sinh(\tfrac{\lambda}{N})+\lambda\bigr]\rho\bigl[r'(\cdot,\widetilde{\theta}\,)\bigr]\\ +\mathcal{K}\bigl[\rho,\pi_{\exp\{-[N\sinh(\frac{\lambda}{N})-\beta]r\}}\bigr]\\ -\log\Bigl\{\pi\Bigl[\exp\bigl\{-\bigl[N\sinh(\tfrac{\lambda}{N})-\beta\bigr]r'(\cdot,\widetilde{\theta}\,)\bigr\}\Bigr]\Bigr\}.\end{aligned}$$

This ends to prove

Theorem 1.31. *For any positive real parameters β and λ, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned}\mathbb{P}\bigl\{\rho\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\}\leq\mathbb{P}\biggl\{\biggl[1-\frac{N\sinh(\frac{\lambda}{N})-\lambda}{\beta}\biggr]\rho\bigl[r'(\cdot,\widetilde{\theta}\,)\bigr]\\ +\frac{\mathcal{K}\bigl[\rho,\pi_{\exp\{-[N\sinh(\frac{\lambda}{N})-\beta]r\}}\bigr]}{\beta}\\ +\beta^{-1}\log\Bigl\{\pi_{\exp\{-[N\sinh(\frac{\lambda}{N})-\beta]r\}}\Bigl[\exp\bigl[N\sinh(\tfrac{\lambda}{N})\tanh(\tfrac{\lambda}{2N})m'(\cdot,\widetilde{\theta}\,)\bigr]\Bigr]\Bigr\}\biggr\}.\end{aligned}$$

Taking $\beta=\frac{N}{2}\sinh(\frac{\lambda}{N})$, using the fact that $\sinh(a)\geq a$, $a\geq 0$ and expressing $\tanh(\frac{a}{2})=a^{-1}\bigl[\sqrt{1+\sinh(a)^2}-1\bigr]$ and $a=\log\bigl[\sqrt{1+\sinh(a)^2}+\sinh(a)\bigr]$, we deduce

Corollary 1.32. *For any positive real constant β and any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\bigl\{\rho\bigl[R'(\cdot,\widetilde{\theta}\,)\bigr]\bigr\}\leq\mathbb{P}\biggl\{\underbrace{\biggl[\frac{N}{\beta}\log\biggl(\sqrt{1+\frac{4\beta^2}{N^2}}+\frac{2\beta}{N}\biggr)-1\biggr]}_{\leq 1}\rho\bigl[r'(\cdot,\widetilde{\theta}\,)\bigr]$$

$$+\frac{1}{\beta}\biggl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\beta r)}\bigr] \\ +\log\biggl[\pi_{\exp(-\beta r)}\Bigl\{\exp\Bigl[N\Bigl(\sqrt{1+\tfrac{4\beta^2}{N^2}}-1\Bigr)m'(\cdot,\widetilde{\theta}\,)\Bigr]\Bigr\}\biggr]\biggr\}\biggr\}.$$

This theorem and its corollary are really anologous to Theorem 1.27 (page 46) and it could easily be proved that under Mammen and Tsybakov margin assumptions, we obtain an upper bound of the same order as Corollary 1.30 (page 48). Anyhow, in order to obtain an empirical bound, we are going now to take a supremum over all possible values of $\widetilde{\theta}$, that is over Θ_1. Although we believe that taking this supremum will not spoil the bound in cases when overfitting remains under control, we will not try to investigate precisely if and when this is actually true, and provide our empirical bound as such. Let us only say that on a qualitative ground, the values of the margin function quantify how steep is the contrast function R or its empirical counterpart r, and that the definition of the empirical margin function is obtained by substituting $\mathbb{P}$, the true sample distribution, with $\overline{\mathbb{P}} = \bigl(\frac{1}{N}\sum_{i=1}^N \delta_{(X_i,Y_i)}\bigr)^{\otimes N}$, the empirical sample distribution, in the definition of the expected margin function. Therefore, on qualitative grounds, it sounds like hopeless to presume that R is steep when r is not, or in other words that a classification model that would be unefficient at estimating a bootstrapped sample according to our non random bound would be by some miracle efficient at estimating the true sample distribution according to the same bound. To this extent, we feel that our empirical bounds bring a satisfactory counterpart of our non random bounds. Anyhow, we will also produce estimators which can be proved to be adaptive using PAC-Bayesian tools in the next subsection, at the price of a more sophisticated construction involving comparisons between a posterior distribution and a Gibbs prior distribution.

Let us restrict now to the important case when $\widetilde{\theta} \in \arg\min_{\Theta_1} R$. To obtain an observable bound, let $\widehat{\theta} \in \arg\min_{\theta\in\Theta} r(\theta)$ and let us introduce the *empirical margin functions*

$$\overline{\varphi}(x) = \sup_{\theta\in\Theta} m'(\theta,\widehat{\theta}) - x\bigl[r(\theta)-r(\widehat{\theta})\bigr], \quad x\in\mathbb{R}_+,$$
$$\widetilde{\varphi}(x) = \sup_{\theta\in\Theta_1} m'(\theta,\widehat{\theta}) - x\bigl[r(\theta)-r(\widehat{\theta})\bigr], \quad x\in\mathbb{R}_+.$$

Using the fact that $m'(\theta,\widetilde{\theta}) \leq m'(\theta,\widehat{\theta}) + m'(\widehat{\theta},\widetilde{\theta})$, we get

COROLLARY 1.33. *For any positive real parameters β and λ, for any posterior distribution $\rho:\Omega\to\mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\bigl[\rho(R)\bigr] - \inf_{\Theta_1} R \leq \mathbb{P}\Biggl\{ \Bigl[1 - \tfrac{N \sinh(\frac{\lambda}{N}) - \lambda}{\beta} \Bigr] \bigl[\rho(r) - r(\widehat{\theta})\bigr] + \frac{\mathcal{K}\bigl[\rho, \pi_{\exp\{-[N\sinh(\frac{\lambda}{N}) - \beta] r\}}\bigr]}{\beta}$$
$$+ \beta^{-1} \log \Bigl\{ \pi_{\exp\{-[N \sinh(\frac{\lambda}{N}) - \beta] r\}} \Bigl[\exp\bigl[N \sinh\bigl(\tfrac{\lambda}{N}\bigr) \tanh\bigl(\tfrac{\lambda}{2N}\bigr) m'(\cdot, \widehat{\theta})\bigr]\Bigr]\Bigr\}$$
$$+ \beta^{-1} N \sinh(\tfrac{\lambda}{N}) \tanh(\tfrac{\lambda}{2N}) \widetilde{\varphi} \Biggl[\frac{\beta}{N \sinh(\frac{\lambda}{N}) \tanh(\frac{\lambda}{2N})} \left(1 - \frac{N \sinh(\frac{\lambda}{N}) - \lambda}{\beta} \right) \Biggr] \Biggr\}.$$

Taking $\beta = \frac{N}{2} \sinh(\frac{\lambda}{N})$, *we also obtain*

$$\mathbb{P}\bigl[\rho(R)\bigr] - \inf_{\Theta_1} R \leq \mathbb{P}\Biggl\{ \underbrace{\Bigl[\tfrac{N}{\beta} \log\Bigl(\sqrt{1 + \tfrac{4\beta^2}{N^2}} + \tfrac{2\beta}{N} \Bigr) - 1 \Bigr]}_{\leq 1} \bigl[\rho(r) - r(\widehat{\theta})\bigr]$$
$$+ \frac{1}{\beta} \Biggl\{ \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr] + \log \Bigl[\pi_{\exp(-\beta r)} \Bigl\{ \exp\Bigl[N \Bigl(\sqrt{1 + \tfrac{4\beta^2}{N^2}} - 1 \Bigr) m'(\cdot, \widehat{\theta}) \Bigr] \Bigr\} \Bigr] \Biggr\}$$
$$+ \frac{N}{\beta} \Bigl(\sqrt{1 + \tfrac{4\beta^2}{N^2}} - 1 \Bigr) \widetilde{\varphi} \left[\frac{\log\Bigl(\sqrt{1 + \frac{4\beta^2}{N^2}} + \frac{2\beta}{N} \Bigr) - \frac{\beta}{N}}{\Bigl(\sqrt{1 + \frac{4\beta^2}{N^2}} - 1 \Bigr)} \right] \Biggr\}.$$

Note that we could also use the upper bound $m'(\theta, \widehat{\theta}) \leq x\bigl[r(\theta) - r(\widehat{\theta})\bigr] + \overline{\varphi}(x)$ and put $\alpha = N \sinh(\frac{\lambda}{N}) \bigl[1 - x \tanh(\frac{\lambda}{2N}) \bigr] - \beta$, to obtain

COROLLARY 1.34. *For any non negative real parameters x, α and λ, such that $\alpha < N \sinh(\frac{\lambda}{N}) \bigl[1 - x \tanh(\frac{\lambda}{2N})\bigr]$, for any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\bigl[\rho(R)\bigr] - \inf_{\Theta_1} R$$
$$\leq \mathbb{P}\Biggl\{ \left[1 - \frac{N \sinh(\frac{\lambda}{N}) \bigl[1 - x \tanh(\frac{\lambda}{2N})\bigr] - \lambda}{N \sinh(\frac{\lambda}{N}) \bigl[1 - x \tanh(\frac{\lambda}{2N})\bigr] - \alpha} \right] \bigl[\rho(r) - r(\widehat{\theta})\bigr]$$
$$+ \frac{\mathcal{K}\bigl[\rho, \pi_{\exp(-\alpha r)}\bigr]}{N \sinh(\frac{\lambda}{N}) \bigl[1 - x \tanh(\frac{\lambda}{2N})\bigr] - \alpha}$$
$$+ \frac{N \sinh(\frac{\lambda}{N}) \tanh(\frac{\lambda}{2N})}{N \sinh(\frac{\lambda}{N}) \bigl[1 - x \tanh(\frac{\lambda}{2N})\bigr] - \alpha}$$

$$\times \left[\overline{\varphi}(x) + \widetilde{\varphi}\left(\frac{\lambda - \alpha}{N \sinh(\frac{\lambda}{N}) \tanh(\frac{\lambda}{2N})} \right) \right] \Bigg\}.$$

Let us notice that in the case when $\Theta_1 = \Theta$, the upper bound provided by this corollary has the same general form as the upper bound provided by Corollary 1.28 (page 46), with the sample distribution $\mathbb{P}$ replaced with the empirical distribution of the sample $\overline{\mathbb{P}} = \bigl(\frac{1}{N} \sum_{i=1}^{N} \delta_{(X_i,Y_i)} \bigr)^{\otimes N}$. Therefore, our empirical bound can be of a larger order of magnitude than our non random bound only in the case when our non random bound applied to the bootstrapped sample distribution $\overline{\mathbb{P}}$ would be of a larger order of magnitude than when applied to the true sample distribution $\mathbb{P}$. In other words, we can say that our empirical bound is close to our non random bound in every situation where the bootstrapped sample distribution $\overline{\mathbb{P}}$ is not harder to bound than the true sample distribution $\mathbb{P}$. Although this does not prove that our empirical bound is always of the same order as our non random bound, this is a good qualitative hint that this will be the case in most practical situations of interest, since in situations of "underfitting", if they exist, it is likely that the choice of the classification model is inappropriate to the data and should be modified.

Another reassuring remark is that the empirical margin functions $\overline{\varphi}$ and $\widetilde{\varphi}$ behave well in the case when $\inf_\Theta r = 0$. Indeed in this case $m'(\theta, \widehat{\theta}) = r'(\theta, \widehat{\theta}) = r(\theta)$, $\theta \in \Theta$, and thus $\overline{\varphi}(1) = \widetilde{\varphi}(1) = 0$, and

$$\widetilde{\varphi}(x) \leq -(x-1) \inf\nolimits_{\Theta_1} r, \; x \geq 1.$$

This shows that we recover in this case the same accuracy as with non relative local empirical bounds. Thus the bound of Corollary 1.34 does not collapse in presence of massive overfitting in the larger model, causing $r(\widehat{\theta}) = 0$, which is another hint that this may be an accurate bound in many situations.

1.4.4. Relative empirical deviation bounds. It is natural to make use of Theorem 1.26 on page 44 to obtain empirical deviation bounds, since this theorem provides an empirical variance term.

Theorem 1.26 is written in a way which exploits the fact that ψ_i takes only the three values -1, 0 and +1. However, it will be more convenient for the following computations to use it in its more general form, which only makes use of the fact that $\psi_i \in \; (-1, 1)$. With notations to be explained hereafter, it can indeed also be written as

$$\mathbb{P} \Biggl\{ \exp \Biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \Biggl\{ -N\rho \Bigl\{ \log \Bigl[1 - \lambda P(\psi) \Bigr] \Bigr\}$$

$$+ N\rho\Big\{\overline{P}\Big[\log(1-\lambda\psi)\Big]\Big\} - \mathcal{K}(\rho,\pi)\Big\}\Big]\Big]\Big\} \leq 1. \quad (1.22)$$

We have used the following notations in this inequality. We have put

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} \delta_{(X_i,Y_i)},$$

so that $\overline{P}$ is our notation for the empirical distribution of the process $(X_i, Y_i)_{i=1}^N$. Moreover we have also used

$$P = \mathbb{P}(\overline{P}) = \frac{1}{N}\sum_{i=1}^{N} P_i,$$

where it should be remembered that the joint distribution of the process $(X_i, Y_i)_{i=1}^N$ is $\mathbb{P} = \bigotimes_{i=1}^N P_i$. We have considered $\psi(\theta, \widetilde{\theta})$ as a function defined on $\mathcal{X} \times \mathcal{Y}$,
as $\psi(\theta,\widetilde{\theta})(x,y) = \mathbb{1}\big[y \neq f_\theta(x)\big] - \mathbb{1}\big[y \neq f_{\widetilde{\theta}}(x)\big], \quad (x,y) \in \mathcal{X} \times \mathcal{Y}$
so that it should be understood that

$$\begin{aligned} P(\psi) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}\big[\psi_i(\theta,\widetilde{\theta})\big] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}\Big\{\mathbb{1}\big[Y_i \neq f_\theta(X_i)\big] - \mathbb{1}\big[Y_i \neq f_{\widetilde{\theta}}(X_i)\big]\Big\} = R'(\theta,\widetilde{\theta}). \end{aligned}$$

In the same way

$$\overline{P}\Big[\log(1-\lambda\psi)\Big] = \frac{1}{N}\sum_{i=1}^{N} \log\big[1 - \lambda\psi_i(\theta,\widetilde{\theta})\big].$$

Moreover integration with respect to ρ bears on the index θ, so that

$$\rho\Big\{\log\Big[1-\lambda P(\psi)\Big]\Big\} = \int_{\theta\in\Theta} \log\Big\{1 - \frac{\lambda}{N}\sum_{i=1}^{N}\mathbb{P}\big[\psi_i(\theta,\widetilde{\theta})\big]\Big\}\rho(d\theta),$$

$$\rho\Big\{\overline{P}\Big[\log(1-\lambda\psi)\Big]\Big\} = \int_{\theta\in\Theta}\Big\{\frac{1}{N}\sum_{i=1}^{N}\log\big[1-\lambda\psi_i(\theta,\widetilde{\theta})\big]\Big\}\rho(d\theta).$$

We have chosen concise notations, as we did throughout these notes, in order to make the computations easier to follow.

To get an alternate version of empirical relative deviation bounds, we need to find some convenient way to localize the choice of the prior distribution π in equation (1.22, page 54). Here we propose to replace π with $\mu = \pi_{\exp\{-N\log[1+\beta P(\psi)]\}}$, which can also be written $\pi_{\exp\{-N\log[1+\beta R'(\cdot,\widetilde{\theta})]\}}$. Indeed we see that

$$\mathcal{K}(\rho,\mu) = N\rho\Bigl\{\log\bigl[1+\beta P(\psi)\bigr]\Bigr\} + \mathcal{K}(\rho,\pi) \\ + \log\Bigl\{\pi\Bigl[\exp\bigl\{-N\log\bigl[1+\beta P(\psi)\bigr]\bigr\}\Bigr]\Bigr\}.$$

Moreover, we deduce from our deviation inequality applied to $-\psi$, that (as long as $\beta > -1$),

$$\mathbb{P}\biggl\{\exp\biggl[N\mu\Bigl\{\overline{P}\bigl[\log(1+\beta\psi)\bigr]\Bigr\} - N\mu\Bigl\{\log\bigl[1+\beta P(\psi)\bigr]\Bigr\}\biggr]\biggr\} \leq 1.$$

Thus

$$\mathbb{P}\biggl\{\exp\biggl[\log\Bigl\{\pi\Bigl[\exp\bigl\{-N\log\bigl[1+\beta P(\psi)\bigr]\bigr\}\Bigr]\Bigr\} \\ - \log\Bigl\{\pi\Bigl[\exp\bigl\{-N\overline{P}\bigl[\log(1+\beta\psi)\bigr]\bigr\}\Bigr]\Bigr\}\biggr]\biggr\} \\ \leq \mathbb{P}\biggl\{\exp\biggl[-N\mu\Bigl\{\log\bigl[1+\beta P(\psi)\bigr]\Bigr\} - \mathcal{K}(\mu,\pi) \\ + N\mu\Bigl\{\overline{P}\bigl[\log(1+\beta\psi)\bigr]\Bigr\} + \mathcal{K}(\mu,\pi)\biggr]\biggr\} \leq 1.$$

This can be used to handle $\mathcal{K}(\rho,\mu)$, making use of the Cauchy Schwarz inequality as follows

$$\mathbb{P}\Biggl\{\exp\Biggl[\frac{1}{2}\biggl[-N\log\Bigl\{\Bigl(1-\lambda\rho\bigl[P(\psi)\bigr]\Bigr)\Bigl(1+\beta\rho\bigl[P(\psi)\bigr]\Bigr)\Bigr\} \\ +N\rho\Bigl\{\overline{P}\Bigl[\log(1-\lambda\psi)\Bigr]\Bigr\} \\ - \mathcal{K}(\rho,\pi) - \log\Bigl\{\pi\Bigl[\exp\bigl\{-N\overline{P}\bigl[\log(1+\beta\psi)\bigr]\bigr\}\Bigr]\Bigr\}\biggr]\Biggr]\Biggr\} \\ \leq \mathbb{P}\biggl\{\exp\biggl[-N\log\Bigl\{\Bigl(1-\lambda\rho\bigl[P(\psi)\bigr]\Bigr)\Bigr\} \\ + N\rho\Bigl\{\overline{P}\Bigl[\log(1-\lambda\psi)\Bigr]\Bigr\} - \mathcal{K}(\rho,\mu)\biggr]\biggr\}^{1/2}$$

$$\times \mathbb{P}\biggl\{ \exp \biggl[\log \Bigl\{ \pi \Bigl[\exp \bigl\{ -N \log \bigl[1 + \beta P(\psi) \bigr] \bigr\} \Bigr] \Bigr\} \\ - \log \Bigl\{ \pi \Bigl[\exp \bigl\{ -N \overline{P} \bigl[\log(1 + \beta\psi) \bigr] \bigr\} \Bigr] \Bigr\} \biggr] \biggr\}^{1/2} \leq 1.$$

This implies that with $\mathbb{P}$ probability at least $1 - \epsilon$,

$$- N \log \Bigl\{ \Bigl(1 - \lambda \rho \bigl[P(\psi) \bigr] \Bigr) \Bigl(1 + \beta \rho \bigl[P(\psi) \bigr] \Bigr) \Bigr\} \\ \leq -N \rho \Bigl\{ \overline{P} \Bigl[\log(1 - \lambda\psi) \Bigr] \Bigr\} \\ + \mathcal{K}(\rho, \pi) + \log \Bigl\{ \pi \Bigl[\exp \bigl\{ -N \overline{P} \bigl[\log(1 + \beta\psi) \bigr] \bigr\} \Bigr] \Bigr\} - 2 \log(\epsilon).$$

It is now convenient to remember that

$$\overline{P} \Bigl[\log(1 - \lambda\psi) \Bigr] = \frac{1}{2} \log \left(\frac{1 - \lambda}{1 + \lambda} \right) r'(\theta, \widetilde{\theta}) + \frac{1}{2} \log(1 - \lambda^2) m'(\theta, \widetilde{\theta}).$$

We thus can write the previous inequality as

$$- N \log \Bigl\{ \Bigl(1 - \lambda \rho \bigl[R'(\cdot, \widetilde{\theta}) \bigr] \Bigr) \Bigl(1 + \beta \rho \bigl[R'(\cdot, \widetilde{\theta}) \bigr] \Bigr) \Bigr\} \\ \leq \frac{N}{2} \log \left(\frac{1 + \lambda}{1 - \lambda} \right) \rho \bigl[r'(\cdot, \widetilde{\theta}) \bigr] - \frac{N}{2} \log(1 - \lambda^2) \rho \bigl[m'(\cdot, \widetilde{\theta}) \bigr] + \mathcal{K}(\rho, \pi) \\ + \log \Biggl\{ \pi \Biggl[\exp \biggl\{ - \frac{N}{2} \log \Bigl(\frac{1 + \beta}{1 - \beta} \Bigr) r'(\cdot, \widetilde{\theta}) \\ - \frac{N}{2} \log(1 - \beta^2) m'(\cdot, \widetilde{\theta}) \biggr\} \Biggr] \Biggr\} - 2 \log(\epsilon).$$

Let us assume now that $\widetilde{\theta} \in \arg\min_{\Theta_1} R$. Let us introduce $\widehat{\theta} \in \arg\min_{\Theta} r$. Decomposing $r'(\theta, \widetilde{\theta}) = r'(\theta, \widehat{\theta}) + r'(\widehat{\theta}, \widetilde{\theta})$ and considering that

$$m'(\theta, \widetilde{\theta}) \leq m'(\theta, \widehat{\theta}) + m'(\widehat{\theta}, \widetilde{\theta}),$$

we see that with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$
\begin{aligned}
- N \log\Bigl\{ \Bigl(1 - \lambda \rho\bigl[R'(\cdot, \widetilde{\theta})\bigr]\Bigr)\Bigl(1 + \beta\rho\bigl[R'(\cdot,\widetilde{\theta})\Bigr)\Bigr\} \\
\leq \frac{N}{2}\log\biggl(\frac{1+\lambda}{1-\lambda}\biggr)\rho\bigl[r'(\cdot,\widehat{\theta})\bigr] - \frac{N}{2}\log(1-\lambda^2)\rho\bigl[m'(\cdot,\widehat{\theta})\bigr] + \mathcal{K}(\rho,\pi) \\
+ \log\biggl\{\pi\biggl[\exp\Bigl\{-\tfrac{N}{2}\log\Bigl(\tfrac{1+\beta}{1-\beta}\Bigr)\bigl[r'(\cdot,\widehat{\theta})\bigr] - \tfrac{N}{2}\log(1-\beta^2)m'(\cdot,\widehat{\theta})\Bigr\}\biggr]\biggr\} \\
+ \tfrac{N}{2}\log\Bigl[\tfrac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\Bigr]\bigl[r(\widehat{\theta}) - r(\widetilde{\theta})\bigr] \\
- \tfrac{N}{2}\log\bigl[(1-\lambda^2)(1-\beta^2)\bigr]m'(\widehat{\theta},\widetilde{\theta}) - 2\log(\epsilon).
\end{aligned}
$$

Let us now define for simplicity the posterior $\nu : \Omega \to \mathcal{M}_+^1(\Theta)$ by the identity

$$
\frac{d\nu}{d\pi}(\theta) = \frac{\exp\Bigl\{-\tfrac{N}{2}\log\Bigl(\tfrac{1+\lambda}{1-\lambda}\Bigr)r'(\theta,\widehat{\theta}) + \tfrac{N}{2}\log(1-\lambda^2)m'(\theta,\widehat{\theta})\Bigr\}}{\pi\biggl[\exp\Bigl\{-\tfrac{N}{2}\log\Bigl(\tfrac{1+\lambda}{1-\lambda}\Bigr)r'(\cdot,\widehat{\theta}) + \tfrac{N}{2}\log(1-\lambda^2)m'(\cdot,\widehat{\theta})\Bigr\}\biggr]}.
$$

Let us also introduce the random bound

$$
\begin{aligned}
B = \frac{1}{N}\log\biggl\{\nu\biggl[\exp\Bigl[\tfrac{N}{2}\log\Bigl[\tfrac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\Bigr]r'(\cdot,\widehat{\theta}) \\
- \tfrac{N}{2}\log\bigl[(1-\lambda^2)(1-\beta^2)\bigr]m'(\cdot,\widehat{\theta})\Bigr]\biggr]\biggr\} \\
+ \sup_{\theta\in\Theta_1}\frac{1}{2}\log\Bigl[\tfrac{(1-\lambda)(1+\beta)}{(1+\lambda)(1-\beta)}\Bigr]r'(\theta,\widehat{\theta}) \\
- \frac{1}{2}\log\bigl[(1-\lambda^2)(1-\beta^2)\bigr]m'(\theta,\widehat{\theta}) - \frac{2}{N}\log(\epsilon).
\end{aligned}
$$

Theorem 1.35. *Using the above notations, for any real constants $0 \leq \beta < \lambda < 1$, for any prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$, for any subset $\Theta_1 \subset \Theta$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$
-\log\Bigl\{\Bigl(1-\lambda\bigl[\rho(R) - \inf_{\Theta_1} R\bigr]\Bigr)\Bigl(1+\beta\bigl[\rho(R) - \inf_{\Theta_1} R\bigr]\Bigr)\Bigr\} \leq \frac{\mathcal{K}(\rho,\nu)}{N} + B.
$$

Therefore,

$$\begin{aligned}\rho(R) - \inf_{\Theta_1} R \\ \leq \frac{\lambda - \beta}{2 \lambda \beta} \left(\sqrt{1 + 4 \frac{\lambda \beta}{(\lambda - \beta)^2} \left[1 - \exp \left(- B - \frac{\mathcal{K}(\rho, \nu)}{N} \right) \right]} - 1 \right) \\ \leq \frac{1}{\lambda - \beta} \left(B + \frac{\mathcal{K}(\rho, \nu)}{N} \right).\end{aligned}$$

Let us define the posterior $\widehat{\nu}$ by the identity

$$\frac{d\widehat{\nu}}{d\pi}(\theta) = \frac{\exp \Bigl[- \frac{N}{2} \log \Bigl(\frac{1+\beta}{1-\beta} \Bigr) r'(\theta, \widehat{\theta}) - \frac{N}{2} \log(1-\beta^2) m'(\theta, \widehat{\theta}) \Bigr]}{\pi \Bigl\{ \exp \Bigl[- \frac{N}{2} \log \Bigl(\frac{1+\beta}{1-\beta} \Bigr) r'(\cdot, \widehat{\theta}) - \frac{N}{2} \log(1-\beta^2) m'(\cdot, \widehat{\theta}) \Bigr] \Bigr\}}.$$

It is useful to remark that

$$\begin{aligned}\frac{1}{N} \log \biggl\{ \nu \biggl[\exp \Bigl[\frac{N}{2} \log \Bigl(\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)} \Bigr) r'(\cdot, \widehat{\theta}) \\ - \frac{N}{2} \log \bigl[(1-\lambda^2)(1-\beta^2) \bigr] m'(\cdot, \widehat{\theta}) \Bigr] \biggr] \biggr\} \\ \leq \widehat{\nu} \biggl\{ \frac{1}{2} \log \Bigl(\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)} \Bigr) r'(\cdot, \widehat{\theta}) \\ - \frac{1}{2} \log \bigl[(1-\lambda^2)(1-\beta^2) \bigr] m'(\cdot, \widehat{\theta}) \biggr\}.\end{aligned}$$

Let us introduce as previously $\overline{\varphi}(x) = \sup_{\theta \in \Theta} m'(\theta, \widehat{\theta}) - x \, r'(\theta, \widehat{\theta})$, $x \in \mathbb{R}_+$. Let us moreover consider $\widetilde{\varphi}(x) = \sup_{\theta \in \Theta_1} m'(\theta, \widehat{\theta}) - x \, r'(\theta, \widehat{\theta})$, $x \in \mathbb{R}_+$. These functions can be used to produce a result which is slightly weaker, but maybe easier to read and understand. Indeed, comming back a little while, we see that, for any $x \in \mathbb{R}_+$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distribution ρ,

$$
\begin{aligned}
- N \log\Bigl\{ \Bigl(1 - \lambda \rho\bigl[R'(\cdot, \widetilde{\theta})\bigr]\Bigr) \Bigl(1 + \beta \rho\bigl[R'(\cdot, \widetilde{\theta})\bigr]\Bigr)\Bigr\} \\
\leq \frac{N}{2} \log\left[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}\right] \rho\bigl[r'(\cdot, \widehat{\theta})\bigr] \\
- \frac{N}{2} \log\bigl[(1-\lambda^2)(1-\beta^2)\bigr] \overline{\varphi}(x) + \mathcal{K}(\rho, \pi) \\
+ \log\biggl\{ \pi \biggl[\exp\Bigl\{ - \tfrac{N}{2} \log\Bigl[\tfrac{(1+\beta)}{(1-\beta)(1-\beta^2)^x}\Bigr] r'(\cdot, \widehat{\theta})\Bigr\}\biggr]\biggr\} \\
- \frac{N}{2} \log\bigl[(1-\lambda^2)(1-\beta^2)\bigr] \widetilde{\varphi}\left(\frac{\log\left[\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\right]}{-\log\left[(1-\lambda^2)(1-\beta^2)\right]}\right) \\
- 2 \log(\epsilon) \\
= \int_{\frac{N}{2}\log\left[\frac{(1+\beta)}{(1-\beta)(1-\beta^2)^x}\right]}^{\frac{N}{2}\log\left[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}\right]} \pi_{\exp(-\alpha r)}\bigl[r'(\cdot, \widehat{\theta})\bigr] d\alpha \\
+ \mathcal{K}\bigl(\rho, \pi_{\exp\{-\frac{N}{2}\log[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}] r\}}\bigr) - 2\log(\epsilon) \\
- \frac{N}{2} \log\bigl[(1-\lambda^2)(1-\beta^2)\bigr] \left[\overline{\varphi}(x) + \widetilde{\varphi}\left(\frac{\log\left[\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\right]}{-\log[(1-\lambda^2)(1-\beta^2)]}\right)\right].
\end{aligned}
$$

Theorem 1.36. *With the previous notations, for any real constants $0 \leq \beta < \lambda < 1$, for any positive real constant x, for any prior probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$, for any subset $\Theta_1 \subset \Theta$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, putting*

$$
\begin{aligned}
B(\rho) = \frac{1}{N(\lambda - \beta)} \int_{\frac{N}{2}\log\left[\frac{(1+\beta)}{(1-\beta)(1-\beta^2)^x}\right]}^{\frac{N}{2}\log\left[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}\right]} \pi_{\exp(-\alpha r)}\bigl[r'(\cdot, \widehat{\theta})\bigr] d\alpha \\
+ \frac{\mathcal{K}\bigl(\rho, \pi_{\exp\{-\frac{N}{2}\log[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}] r\}}\bigr) - 2\log(\epsilon)}{N(\lambda-\beta)} \\
- \frac{1}{2(\lambda-\beta)} \log\bigl[(1-\lambda^2)(1-\beta^2)\bigr] \left[\overline{\varphi}(x) + \widetilde{\varphi}\left(\frac{\log\left[\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\right]}{-\log[(1-\lambda^2)(1-\beta^2)]}\right)\right] \\
\leq \frac{1}{N(\lambda-\beta)} d_e \log\left(\frac{\log\Bigl[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}\Bigr]}{\log\Bigl(\frac{(1+\beta)}{(1-\beta)(1-\beta^2)^x}\Bigr)}\right)
\end{aligned}
$$

$$+\frac{\mathcal{K}(\rho,\pi_{\exp\{-\frac{N}{2}\log[\frac{(1+\lambda)}{(1-\lambda)(1-\lambda^2)^x}]r\}})-2\log(\epsilon)}{N(\lambda-\beta)}$$
$$-\frac{1}{2(\lambda-\beta)}\log\bigl[(1-\lambda^2)(1-\beta^2)\bigr]\left[\overline{\varphi}(x)+\widetilde{\varphi}\left(\frac{\log\Bigl[\frac{(1+\lambda)(1-\beta)}{(1-\lambda)(1+\beta)}\Bigr]}{-\log[(1-\lambda^2)(1-\beta^2)]}\right)\right],$$

the following bounds hold true:

$$\begin{aligned}\rho(R)-\inf_{\Theta_1}R\\ \leq \frac{\lambda-\beta}{2\lambda\beta}\left(\sqrt{1+\frac{4\lambda\beta}{(\lambda-\beta)^2}\Bigl\{1-\exp\bigl[-(\lambda-\beta)B(\rho)\bigr]\Bigr\}}-1\right)\\ \leq B(\rho).\end{aligned}$$

Let us remark that this alternative way of handling relative deviation bounds made it possible to carry on with non linear bounds up to the final result. (For instance, if $\lambda = 0.5$, $\beta = 0.2$ and $B(\rho) = 0.1$, the non linear bound gives $\rho(R) - \inf_{\Theta_1} R \leq 0.096$.)

1.5. BOUNDS RELATIVE TO A GIBBS DISTRIBUTION. The empirical bounds of the previous section involve taking suprema in $\theta \in \Theta$, and replacing the *margin function* φ by some empirical counter parts $\overline{\varphi}$ or $\widetilde{\varphi}$, which may prove unsafe when using very complex classification models. Moreover, they are not easy to analyze with PAC-Bayesian tools. To remedy these weaknesses, we are going now to propose another type of relative bounds. We will first explain how to compare the expected error rate $\rho(R)$ of any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ with $\pi_{\exp(-\beta R)}(R)$, the expected risk of a Gibbs prior distribution. We will then show how to analyze the behaviour of this bound. This will provide an estimator proven to reach adaptively the best possible asymptotic behaviour of the error rate under Mammen and Tsybakov margin assumptions and parametric complexity assumptions.

Then, we will provide an empirical bound for the Kullback divergence $\mathcal{K}(\rho, \pi_{\exp(-\beta R)})$ of a posterior distribution with respect to a Gibbs prior, making use of relative deviation inequalities.

To tackle the question of model selection, we will estimate the relative performance of one posterior distribution with respect to another, which is useful when the two posteriors are supported by different models.

Eventually, we will propose a more integrated approach to model selection, showing how to build a two step localization strategy, in which the

performance of the posterior distribution to be analyzed is compared with some *two step* Gibbs prior.

1.5.1. Comparing a posterior distribution with a Gibbs prior. Similarly to Theorem 1.26 we can prove that for any prior distribution $\widetilde{\pi} \in \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\biggl\{ \widetilde{\pi} \otimes \widetilde{\pi} \biggl\{ \exp \biggl[-N \log(1 - \lambda R') \\ - \frac{N}{2} \log \left(\frac{1+\lambda}{1-\lambda} \right) r' + \frac{N}{2} \log\bigl(1 - \lambda^2\bigr) m' \biggr] \biggr\} \biggr\} \leq 1. \quad (1.23)$$

Replacing $\widetilde{\pi}$ with $\pi_{\exp(-\beta R)}$ and considering the posterior distribution $\rho \otimes \pi_{\exp(-\beta R)}$, provides a starting point in the comparison of ρ with $\pi_{\exp(-\beta R)}$; we can indeed state with $\mathbb{P}$ probability at least $1 - \epsilon$ that

$$\begin{aligned} - N \log \Bigl\{ 1 - \lambda \Bigl[\rho(R) - \pi_{\exp(-\beta R)}(R) \Bigr] \Bigr\} \\ \leq \frac{N}{2} \log \left(\frac{1+\lambda}{1-\lambda} \right) \bigl[\rho(r) - \pi_{\exp(-\beta R)}(r) \bigr] \\ - \frac{N}{2} \log\bigl(1 - \lambda^2\bigr) \rho \otimes \pi_{\exp(-\beta R)}(m') \\ + \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] - \log(\epsilon). \quad (1.24) \end{aligned}$$

Using the parameter $\gamma = \frac{N}{2} \log \left(\frac{1+\lambda}{1-\lambda} \right)$, so that $\lambda = \tanh \left(\frac{\gamma}{N} \right)$ and $-\frac{N}{2} \log(1-\lambda^2) = N \log\bigl[\cosh(\frac{\gamma}{N})\bigr]$, and noticing that

$$\begin{aligned} \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] = \beta \bigl[\rho(R) - \pi_{\exp(-\beta R)}(R) \bigr] \\ + \mathcal{K}(\rho, \pi) - \mathcal{K}\bigl[\pi_{\exp(-\beta R)}, \pi\bigr], \quad (1.25) \end{aligned}$$

makes a step further in the proper handling of the entropy term:

$$\begin{aligned} - N \log \Bigl\{ 1 - \tanh(\tfrac{\gamma}{N}) \Bigl[\rho(R) - \pi_{\exp(-\beta R)}(R) \Bigr] \Bigr\} - \beta \bigl[\rho(R) - \pi_{\exp(-\beta R)}(R) \bigr] \\ \leq \gamma \bigl[\rho(r) - \pi_{\exp(-\beta R)}(r) \bigr] + N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] \rho \otimes \pi_{\exp(-\beta R)}(m') \\ + \mathcal{K}(\rho, \pi) - \mathcal{K}\bigl[\pi_{\exp(-\beta R)}, \pi\bigr] - \log(\epsilon). \quad (1.26) \end{aligned}$$

We can then decompose in the right-hand side $\gamma\bigl[\rho(r) - \pi_{\exp(-\beta R)}(r)\bigr]$ into $(\gamma - \lambda)\bigl[\rho(r) - \pi_{\exp(-\beta R)}(r)\bigr] + \lambda\bigl[\rho(r) - \pi_{\exp(-\beta R)}(r)\bigr]$ and use the fact that

$$
\begin{aligned}
\lambda\bigl[\rho(r) - \pi_{\exp(-\beta R)}(r)\bigr] + N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\, \rho \otimes \pi_{\exp(-\beta R)}(m') \\
+ \mathcal{K}(\rho, \pi) - \mathcal{K}\bigl[\pi_{\exp(-\beta R)}, \pi\bigr] \\
\leq \lambda \rho(r) + \mathcal{K}(\rho, \pi) + \log\Bigl\{ \pi\Bigl[\exp\bigl\{ -\lambda r + N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] \rho(m')\bigr\}\Bigr]\Bigr\} \\
= \mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr] + \log\Bigl\{ \pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{ N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] \rho(m')\bigr\}\Bigr]\Bigr\},
\end{aligned}
$$

to get rid of the appearance of the unobserved Gibbs prior $\pi_{\exp(-\beta R)}$ in most places of the right-hand side of our inequality, leading to

THEOREM 1.37. *For any real constants β and γ, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, for any real constant λ,*

$$
\begin{aligned}
&\bigl[N \tanh(\tfrac{\gamma}{N}) - \beta\bigr]\bigl[\rho(R) - \pi_{\exp(-\beta R)}(R)\bigr] \\
&\quad \leq -N \log\Bigl\{ 1 - \tanh(\tfrac{\gamma}{N}) \Bigl[\rho(R) - \pi_{\exp(-\beta R)}(R)\Bigr]\Bigr\} \\
&\qquad\qquad - \beta\bigl[\rho(R) - \pi_{\exp(-\beta R)}(R)\bigr] \\
&\quad \leq (\gamma - \lambda)\bigl[\rho(r) - \pi_{\exp(-\beta R)}(r)\bigr] + \mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr] \\
&\qquad + \log\Bigl\{ \pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] \rho(m')\bigr\}\Bigr]\Bigr\} - \log(\epsilon) \\
&\quad = \mathcal{K}\bigl[\rho, \pi_{\exp(-\gamma r)}\bigr] \\
&\qquad + \log\Bigl\{ \pi_{\exp(-\gamma r)}\Bigl[\exp\bigl\{(\gamma - \lambda) r + N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] \rho(m')\bigr\}\Bigr]\Bigr\} \\
&\qquad\qquad - (\gamma - \lambda)\pi_{\exp(-\beta R)}(r) - \log(\epsilon).
\end{aligned}
$$

We would like to have a fully empirical upper bound even in the case when $\lambda \neq \gamma$. This can be done by using the theorem twice. We will need a lemma

LEMMA 1.38 *For any probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$, for any bounded measurable functions $g, h : \Theta \to \mathbb{R}$,*

$$
\pi_{\exp(-g)}(g) - \pi_{\exp(-h)}(g) \leq \pi_{\exp(-g)}(h) - \pi_{\exp(-h)}(h).
$$

PROOF. Let us notice that

$$
\begin{aligned}
0 \leq \mathcal{K}(\pi_{\exp(-g)}, \pi_{\exp(-h)}) &= \pi_{\exp(-g)}(h) + \log\bigl\{\pi\bigl[\exp(-h)\bigr]\bigr\} + \mathcal{K}(\pi_{\exp(-g)}, \pi) \\
&= \pi_{\exp(-g)}(h) - \pi_{\exp(-h)}(h) - \mathcal{K}(\pi_{\exp(-h)}, \pi) + \mathcal{K}(\pi_{\exp(-g)}, \pi) \\
&= \pi_{\exp(-g)}(h) - \pi_{\exp(-h)}(h) - \mathcal{K}(\pi_{\exp(-h)}, \pi) - \pi_{\exp(-g)}(g) - \log\bigl\{\pi\bigl[\exp(-g)\bigr]\bigr\}.
\end{aligned}
$$

Moreover

$$-\log\bigl\{\pi\bigl[\exp(-g)\bigr]\bigr\} \leq \pi_{\exp(-h)}(g) + \mathcal{K}(\pi_{\exp(-h)}, \pi),$$

which achieves the proof. □

For any positive real constants β and λ, we can then apply Theorem 1.37 to $\rho = \pi_{\exp(-\lambda r)}$, and use the inequality

$$\frac{\lambda}{\beta}\bigl[\pi_{\exp(-\lambda r)}(r) - \pi_{\exp(-\beta R)}(r)\bigr] \leq \pi_{\exp(-\lambda r)}(R) - \pi_{\exp(-\beta R)}(R) \qquad (1.27)$$

provided by the previous lemma. We thus obtain with $\mathbb{P}$ probability at least $1-\epsilon$

$$\begin{aligned}-N\log\Bigl\{1-\tanh(\tfrac{\gamma}{N})\tfrac{\lambda}{\beta}\Bigl[\pi_{\exp(-\lambda r)}(r) - \pi_{\exp(-\beta R)}(r)\Bigr]\Bigr\}& \\ - \gamma\bigl[\pi_{\exp(-\lambda r)}(r) - \pi_{\exp(-\beta R)}(r)\bigr]& \\ \leq \log\Bigl\{\pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\pi_{\exp(-\lambda r)}(m')\bigr\}\Bigr]\Bigr\} - \log(\epsilon).&\end{aligned}$$

Let us introduce the convex function

$$F_{\gamma,\alpha}(x) = -N\log\bigl[1-\tanh(\tfrac{\gamma}{N})x\bigr] - \alpha x \geq \bigl[N\tanh(\tfrac{\gamma}{N}) - \alpha\bigr]x.$$

With $\mathbb{P}$ probability at least $1-\epsilon$,

$$\begin{aligned}-\pi_{\exp(-\beta R)}(r) \leq \inf_{\lambda\in\mathbb{R}_+^*}\Biggl\{&-\pi_{\exp(-\lambda r)}(r) \\ &+ \frac{\beta}{\lambda}F^{-1}_{\gamma,\frac{\beta\gamma}{\lambda}}\Biggl[\log\Bigl\{\pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\pi_{\exp(-\lambda r)}(m')\bigr\}\Bigr]\Bigr\} \\ &\qquad - \log(\epsilon)\Biggr]\Biggr\}.\end{aligned}$$

Since Theorem 1.37 holds uniformly for any posterior distribution ρ, we can apply it again to some arbitrary posterior distribution ρ. We can moreover make the result uniform in β and γ by considering some atomic measure $\nu \in \mathcal{M}_+^1(\mathbb{R})$ on the real line and using a union bound. This leads to

THEOREM 1.39. *For any atomic probability distribution on the positive real line $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$, for any positive real constants β and γ,*

$$
\begin{aligned}
&\bigl[N\tanh(\tfrac{\gamma}{N})-\beta\bigr]\bigl[\rho(R)-\pi_{\exp(-\beta R)}(R)\bigr] \\
&\qquad\qquad \leq F_{\gamma,\beta}\bigl[\rho(R)-\pi_{\exp(-\beta R)}(R)\bigr] \leq B(\rho,\beta,\gamma), \ \textit{where}
\end{aligned}
$$

$$
\begin{aligned}
B(\rho,\beta,\gamma) = & \inf_{\substack{\lambda_1\in\mathbb{R}_+,\lambda_1\leq\gamma\\ \lambda_2\in\mathbb{R},\lambda_2>\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1}}} \Biggl\{ \mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda_1 r)}\bigr] \\
& + (\gamma-\lambda_1)\bigl[\rho(r)-\pi_{\exp(-\lambda_2 r)}(r)\bigr] \\
& + \log\Bigl\{\pi_{\exp(-\lambda_1 r)}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\rho(m')\bigr\}\Bigr]\Bigr\} - \log\bigl[\epsilon\nu(\beta)\nu(\gamma)\bigr] \\
& + (\gamma-\lambda_1)\frac{\beta}{\lambda_2}F^{-1}_{\gamma,\frac{\beta\gamma}{\lambda_2}}\Biggl[\log\Bigl\{ \\
& \qquad \pi_{\exp(-\lambda_2 r)}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\pi_{\exp(-\lambda_2 r)}(m')\bigr\}\Bigr]\Bigr\} \\
& \qquad\qquad\qquad\qquad - \log\bigl[\epsilon\nu(\beta)\nu(\gamma)\bigr]\Biggr]\Biggr\} \\
\leq & \inf_{\substack{\lambda_1\in\mathbb{R}_+,\lambda_1\leq\gamma\\ \lambda_2\in\mathbb{R},\lambda_2>\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1}}} \Biggl\{ \mathcal{K}\bigl[\rho,\pi_{\exp(-\lambda_1 r)}\bigr] \\
& + (\gamma-\lambda_1)\bigl[\rho(r)-\pi_{\exp(-\lambda_2 r)}(r)\bigr] \\
& + \log\Bigl\{\pi_{\exp(-\lambda_1 r)}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\rho(m')\bigr\}\Bigr]\Bigr\} \\
& + \frac{\beta}{\lambda_2}\frac{(1-\frac{\lambda_1}{\gamma})}{\bigl[\frac{N}{\gamma}\tanh(\frac{\gamma}{N})-\frac{\beta}{\lambda_2}\bigr]}\log\Bigl\{\pi_{\exp(-\lambda_2 r)}\Bigl[\\
& \qquad \exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\pi_{\exp(-\lambda_2 r)}(m')\bigr\}\Bigr]\Bigr\} \\
& \qquad - \Biggl\{1+\frac{\beta}{\lambda_2}\frac{(1-\frac{\lambda_1}{\gamma})}{[\frac{N}{\gamma}\tanh(\frac{\gamma}{N})-\frac{\beta}{\lambda_2}]}\Biggr\}\log\bigl[\epsilon\nu(\beta)\nu(\gamma)\bigr]\Biggr\},
\end{aligned}
$$

where we have written for short $\nu(\beta)$ and $\nu(\gamma)$ instead of $\nu(\{\beta\})$ and $\nu(\{\gamma\})$.

Let us notice that $B(\rho,\beta,\gamma)=+\infty$ when $\nu(\beta)=0$ or $\nu(\gamma)=0$, the uniformity in β and γ of the theorem therefore necessarily bears on a countable number of values of these parameters. We can typically choose for ν distributions such as the one used in Theorem 1.11 on page 21: namely we can put for some positive real ratio $\alpha>1$

$$
\nu(\alpha^k)=\frac{1}{(k+1)(k+2)}, \qquad k\in\mathbb{N},
$$

or alternatively, since we are interested in values of the parameters less than

N, we can prefer

$$\nu(\alpha^k) = \frac{\log(\alpha)}{\log(\alpha N)}, \qquad 0 \leq k < \frac{\log(N)}{\log(\alpha)}.$$

We can also use such a coding distribution on dyadic numbers as the one defined by equation (1.6) on page 23.

1.5.2. The effective temperature of a posterior distribution. Using the parametric approximation $\pi_{\exp(-\alpha r)}(r) - \inf_\Theta r \simeq \frac{d_e}{\alpha}$, we get as an order of magnitude

$$\begin{aligned}
B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma) &\lesssim -(\gamma - \lambda_1) d_e \bigl[\lambda_2^{-1} - \lambda_1^{-1}\bigr] \\
&+ 2 d_e \log \frac{\lambda_1}{\lambda_1 - N \log\bigl[\cosh(\frac{\gamma}{N})\bigr] x} \\
&\quad + 2 \frac{\beta}{\lambda_2} \frac{(1 - \frac{\lambda_1}{\gamma})}{\bigl[\frac{N}{\gamma} \tanh(\frac{\gamma}{N}) - \frac{\beta}{\lambda_2}\bigr]} d_e \log \left(\frac{\lambda_2}{\lambda_2 - N \log\bigl[\cosh(\frac{\gamma}{N})\bigr] x} \right) \\
&\qquad + 2N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] \left[1 + \frac{\beta}{\lambda_2} \frac{(1 - \frac{\lambda_1}{\gamma})}{\bigl[\frac{N}{\gamma} \tanh(\frac{\gamma}{N}) - \frac{\beta}{\lambda_2}\bigr]} \right] \widetilde{\varphi}(x) \\
&\qquad\quad - \left\{ 1 + \frac{\beta}{\lambda_2} \frac{(1 - \frac{\lambda_1}{\gamma})}{\bigl[\frac{N}{\gamma} \tanh(\frac{\gamma}{N}) - \frac{\beta}{\lambda_2}\bigr]} \right\} \log\bigl[\nu(\beta)\nu(\gamma)\epsilon\bigr].
\end{aligned}$$

Therefore, if the empirical dimension d_e stays bounded when N increases, we are going to obtain a negative upper bound for any values of the constants $\lambda_1 > \lambda_2 > \beta$, as soon as γ and $\frac{N}{\gamma}$ are chosen to be large enough. This ability to obtain negative values for the bound $B(\pi_{\exp(-\lambda_1 r)}, \gamma, \beta)$, and more generally $B(\rho, \gamma, \beta)$, leads the way to introducing the new concept of the *effective temperature* of an estimator.

Definition 1.1 For any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ we define the *effective temperature* $T(\rho) \in \mathbb{R} \cup \{-\infty, +\infty\}$ of ρ by the equation

$$\rho(R) = \pi_{\exp(-\frac{R}{T(\rho)})}(R).$$

Note that $\beta \mapsto \pi_{\exp(-\beta R)}(R) : \mathbb{R} \cup \{-\infty, +\infty\} \to (0, 1)$ is continuous and strictly decreasing from $\operatorname{ess\,sup}_\pi R$ to $\operatorname{ess\,inf}_\pi R$ (as soon as these two bounds do not coincide). This shows that the effective temperature $T(\rho)$ is a well defined random variable.

Theorem 1.39 provides a bound for $T(\rho)$, indeed:

PROPOSITION 1.40. *Let*

$$\widehat{\beta}(\rho) = \sup\bigl\{\beta \in \mathbb{R}; \inf_{\gamma, N \tanh(\frac{\gamma}{N}) > \beta} B(\rho, \beta, \gamma) \leq 0\bigr\},$$

where $B(\rho, \beta, \gamma)$ is as in Theorem 1.39. Then with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, $T(\rho) \leq \widehat{\beta}(\rho)^{-1}$, or equivalently $\rho(R) \leq \pi_{\exp[-\widehat{\beta}(\rho)R]}(R)$.

This notion of *effective temperature* of a (randomized) estimator ρ is interesting for two reasons:

- the difference $\rho(R) - \pi_{\exp(-\beta R)}(R)$ can be estimated with a better accuracy than $\rho(R)$ itself, due to the use of relative deviation inequalities, leading to convergence rates up to $1/N$ in favourable situations, even when $\inf_\Theta R$ is not close to zero;
- and of course $\pi_{\exp(-\beta R)}(R)$ is a decreasing function of β, thus being able to estimate $\rho(R) - \pi_{\exp(-\beta R)}(R)$ with some given accuracy, means being able to discriminate between values of $\rho(R)$ with the same accuracy, although doing so through the parametrization $\beta \mapsto \pi_{\exp(-\beta R)}(R)$, which cannot be observed nor estimated with the same precision!

1.5.3. Analysis of an empirical bound for the effective temperature. We are now going to launch into a mathematically rigorous analysis of the bound $B(\pi_{\exp(-\lambda_1 r), \beta, \gamma})$ provided by Theorem 1.39, to show that $\inf_{\rho \in \mathcal{M}_+^1(\Theta)} \pi_{\exp[-\widehat{\beta}(\rho)R]}(R)$ converges indeed to $\inf_\Theta R$ at some unimprovable rates in favourable situations.

It is more convenient for this purpose to use deviation inequalities involving M' rather than m'. It is straightforward to extend Theorem 1.25 on page 43 to

THEOREM 1.41. *For any real constants β and γ, for any prior distribution $\mu \in \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \eta$, for any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\gamma \rho \otimes \pi_{\exp(-\beta R)}\bigl[\Psi_{\frac{\gamma}{N}}(R', M')\bigr] \leq \gamma \rho \otimes \pi_{\exp(-\beta R)}(r') + \mathcal{K}(\rho, \mu) - \log(\eta).$$

In order to transform the left-hand side into a linear expression and in the same time to localize this theorem, let us choose μ defined by its density

$$\frac{d\mu}{d\pi}(\theta_1) = C^{-1} \exp\biggl[-\beta R(\theta_1) \\ - \gamma \int_{\Theta} \Bigl\{ \Psi_{\frac{\gamma}{N}}\bigl[R'(\theta_1,\theta_2), M'(\theta_1,\theta_2)\bigr] \\ - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N}) R'(\theta_1,\theta_2)\Bigr\} \pi_{\exp(-\beta R)}(d\theta_2)\biggr],$$

where C is such that $\mu(\Theta) = 1$. We get

$$\begin{aligned} \mathcal{K}(\rho,\mu) &= \beta\rho(R) + \gamma\rho \otimes \pi_{\exp(-\beta R)}\bigl[\Psi_{\frac{\gamma}{N}}(R',M') - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'\bigr] + \mathcal{K}(\rho,\pi) \\ &\quad + \log\biggl\{ \int_{\Theta} \exp\biggl[-\beta R(\theta_1) \\ &\qquad - \gamma \int_{\Theta}\Bigl\{ \Psi_{\frac{\gamma}{N}}\bigl[R'(\theta_1,\theta_2), M'(\theta_1,\theta_2)\bigr] \\ &\qquad\quad - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'(\theta_1,\theta_2)\Bigr\}\pi_{\exp(-\beta R)}(d\theta_2)\biggr]\pi(d\theta_1)\biggr\} \\ &= \beta\bigl[\rho(R) - \pi_{\exp(-\beta R)}(R)\bigr] \\ &\qquad + \gamma\rho\otimes\pi_{\exp(-\beta R)}\bigl[\Psi_{\frac{\gamma}{N}}(R',M') - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'\bigr] \\ &\qquad\qquad + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi) \\ &\quad + \log\biggl\{\int_{\Theta}\exp\biggl[-\gamma\int_{\Theta}\Bigl\{\Psi_{\frac{\gamma}{N}}\bigl[R'(\theta_1,\theta_2),M'(\theta_1,\theta_2)\bigr] \\ &\qquad - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'(\theta_1,\theta_2)\Bigr\}\pi_{\exp(-\beta R)}(d\theta_2)\biggr]\pi_{\exp(-\beta R)}(d\theta_1)\biggr\}. \end{aligned}$$

Thus with $\mathbb{P}$ probability at least $1-\eta$,

$$\begin{aligned} &\bigl[N\sinh(\tfrac{\gamma}{N}) - \beta\bigr]\bigl[\rho(R) - \pi_{\exp(-\beta R)}(R)\bigr] \\ &\qquad \leq \gamma\bigl[\rho(r) - \pi_{\exp(-\beta R)}(r)\bigr] + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi) - \log(\eta) + C(\beta,\gamma) \\ &\text{where } C(\beta,\gamma) = \log\biggl\{\int_{\Theta}\exp\biggl[-\gamma\int_{\Theta}\Bigl\{\Psi_{\frac{\gamma}{N}}\bigl[R'(\theta_1,\theta_2),M'(\theta_1,\theta_2)\bigr] \\ &\qquad - \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})R'(\theta_1,\theta_2)\Bigr\}\pi_{\exp(-\beta R)}(d\theta_2)\biggr]\pi_{\exp(-\beta R)}(d\theta_1)\biggr\}. \end{aligned} \tag{1.28}$$

Remarking that

$$\mathcal{K}\bigl[\rho,\pi_{\exp(-\beta R)}\bigr] = \beta\bigl[\rho(R) - \pi_{\exp(-\beta R)}(R)\bigr] + \mathcal{K}(\rho,\pi) - \mathcal{K}(\pi_{\exp(-\beta R)},\pi),$$

we deduce from the previous inequality

THEOREM 1.42. *For any real constants β and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$N \sinh(\tfrac{\gamma}{N}) \bigl[\rho(R) - \pi_{\exp(-\beta R)}(R) \bigr] \leq \gamma \bigl[\rho(r) - \pi_{\exp(-\beta R)}(r) \bigr] \\ + \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)} \bigr] - \log(\eta) + C(\beta, \gamma).$$

We can also go into a slightly different direction, starting back again from equation (1.28) on page 67 and remarking that for any real constant λ,

$$\lambda \bigl[\rho(r) - \pi_{\exp(-\beta R)}(r) \bigr] + \mathcal{K}(\rho, \pi) - \mathcal{K}(\pi_{\exp(-\beta R)}, \pi) \\ \leq \lambda \rho(r) + \mathcal{K}(\rho, \pi) + \log \bigl\{ \pi \bigl[\exp(-\lambda r) \bigr] \bigr\} = \mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)} \bigr].$$

This leads to

THEOREM 1.43. *For any real constants β and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any real constant λ,*

$$\bigl[N \sinh(\tfrac{\gamma}{N}) - \beta \bigr] \bigl[\rho(R) - \pi_{\exp(-\beta R)}(R) \bigr] \\ \leq (\gamma - \lambda) \bigl[\rho(r) - \pi_{\exp(-\beta R)}(r) \bigr] + \mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)} \bigr] - \log(\eta) + C(\beta, \gamma),$$

where the definition of $C(\beta, \gamma)$ is given by equation (1.28) *on page 67.*

We can now use this inequality in the case when $\rho = \pi_{\exp(-\lambda r)}$ and combine it with inequality (1.27) on page 63 to obtain

THEOREM 1.44 *For any real constants β and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any real constant λ,*

$$\bigl[\tfrac{N\lambda}{\beta} \sinh(\tfrac{\gamma}{N}) - \gamma \bigr] \bigl[\pi_{\exp(-\lambda r)}(r) - \pi_{\exp(-\beta R)}(r) \bigr] \leq C(\beta, \gamma) - \log(\eta).$$

We deduce from this theorem

PROPOSITION 1.45 *For any real positive constants β_1, β_2 and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any real constants λ_1 and λ_2, such that $\lambda_2 < \beta_2 \frac{\gamma}{N} \sinh(\frac{\gamma}{N})^{-1}$ and $\lambda_1 > \beta_1 \frac{\gamma}{N} \sinh(\frac{\gamma}{N})^{-1}$,*

$$\pi_{\exp(-\lambda_1 r)}(r) - \pi_{\exp(-\lambda_2 r)}(r) \leq \pi_{\exp(-\beta_1 R)}(r) - \pi_{\exp(-\beta_2 R)}(r) \\ + \frac{C(\beta_1, \gamma) + \log(2/\eta)}{\frac{N\lambda_1}{\beta_1} \sinh(\frac{\gamma}{N}) - \gamma} + \frac{C(\beta_2, \gamma) + \log(2/\eta)}{\gamma - \frac{N\lambda_2}{\beta_2} \sinh(\frac{\gamma}{N})}.$$

Moreover, $\pi_{\exp(-\beta_1 R)}$ and $\pi_{\exp(-\beta_2 R)}$ being prior distributions, with $\mathbb{P}$ probability at least $1-\eta$,

$$\gamma\bigl[\pi_{\exp(-\beta_1 R)}(r) - \pi_{\exp(-\beta_2 R)}(r)\bigr] \\ \leq \gamma \pi_{\exp(-\beta_1 R)} \otimes \pi_{\exp(-\beta_2 R)}\bigl[\Psi_{-\frac{\gamma}{N}}(R', M')\bigr] - \log(\eta).$$

Hence

PROPOSITION 1.46 *For any positive real constants β_1, β_2 and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any positive real constants λ_1 and λ_2 such that $\lambda_2 < \beta_2 \frac{\gamma}{N}\sinh(\frac{\gamma}{N})^{-1}$ and $\lambda_1 > \beta_1 \frac{\gamma}{N}\sinh(\frac{\gamma}{N})^{-1}$,*

$$\pi_{\exp(-\lambda_1 r)}(r) - \pi_{\exp(-\lambda_2 r)}(r) \\ \leq \pi_{\exp(-\beta_1 R)} \otimes \pi_{\exp(-\beta_2 R)}\bigl[\Psi_{-\frac{\gamma}{N}}(R', M')\bigr] \\ + \frac{\log(\frac{3}{\eta})}{\gamma} + \frac{C(\beta_1,\gamma) + \log(\frac{3}{\eta})}{\frac{N\lambda_1}{\beta_1}\sinh(\frac{\gamma}{N}) - \gamma} + \frac{C(\beta_2,\gamma) + \log(\frac{3}{\eta})}{\gamma - \frac{N\lambda_2}{\beta_2}\sinh(\frac{\gamma}{N})}.$$

In order to achieve the analysis of the bound $B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma)$ given by Theorem 1.39, there remains now to bound quantities of the general form

$$\log\Bigl\{\pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\pi_{\exp(-\lambda r)}(m')\bigr\}\Bigr]\Bigr\} \\ = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\rho \otimes \pi_{\exp(-\lambda)}(m') - \mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr].$$

Let us consider the prior distribution $\mu \in \mathcal{M}_+^1(\Theta \times \Theta)$ on couples of parameters defined by its density

$$\frac{d\mu}{d(\pi \otimes \pi)}(\theta_1, \theta_2) = C^{-1}\exp\Bigl\{-\beta R(\theta_1) - \beta R(\theta_2) + \alpha \Phi_{-\frac{\alpha}{N}}\bigl[M'(\theta_1, \theta_2)\bigr]\Bigr\},$$

where the normalizing constant C is such that $\mu(\Theta \times \Theta) = 1$. Since for fixed values of the parameters θ and $\theta' \in \Theta$, $m'(\theta, \theta')$, like $r(\theta)$, is a sum of independent Bernoulli random variables, we can easily adapt the proof of Theorem 1.4 on page 11, to establish that with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution ρ and any real constant λ,

$$\alpha \rho \otimes \pi_{\exp(-\lambda r)}(m') \leq \alpha \rho \otimes \pi_{\exp(-\lambda r)}\bigl[\Phi_{-\frac{\alpha}{N}}(M')\bigr] \\ + \mathcal{K}(\rho \otimes \pi_{\exp(-\lambda r)}, \mu) - \log(\eta)$$

$$= \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] + \mathcal{K}\bigl[\pi_{\exp(-\lambda r)}, \pi_{\exp(-\beta R)}\bigr] \\ + \log\Bigl\{\pi_{\exp(-\beta R)} \otimes \pi_{\exp(-\beta R)}\Bigl[\exp\bigl(\alpha\Phi_{-\frac{\alpha}{N}} \circ M'\bigr)\Bigr]\Bigr\} - \log(\eta).$$

Thus for any real constant β and any positive real constants α and γ, with $\mathbb{P}$ probability at least $1-\eta$, for any real constant λ,

$$\log\Bigl\{\pi_{\exp(-\lambda r)}\Bigl[\exp\bigl\{N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\pi_{\exp(-\lambda r)}(m')\bigr\}\Bigr]\Bigr\} \\ \leq \sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Biggl(\frac{N}{\alpha}\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\Bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] + \mathcal{K}\bigl[\pi_{\exp(-\lambda r)}, \pi_{\exp(-\beta R)}\bigr] \\ + \log\bigl\{\pi_{\exp(-\beta R)} \otimes \pi_{\exp(-\beta R)}\bigl[\exp(\alpha\Phi_{-\frac{\alpha}{N}} \circ M')\bigr]\bigr\} \\ - \log(\eta)\Bigr\} - \mathcal{K}\bigl[\rho, \pi_{\exp(-\lambda r)}\bigr]\Biggr). \quad (1.29)$$

To conclude, we need some suitable upper bound for the entropy $\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr]$. This question can be handled in the following way: using Theorem 1.42 on page 68, we see that for any positive real constants γ and β, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution ρ,

$$\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] = \beta\bigl[\rho(R) - \pi_{\exp(-\beta R)}(R)\bigr] + \mathcal{K}(\rho, \pi) - \mathcal{K}(\pi_{\exp(-\beta R)}, \pi) \\ \leq \frac{\beta}{N\sinh(\frac{\gamma}{N})}\Bigl[\gamma\bigl[\rho(r) - \pi_{\exp(-\beta R)}(r)\bigr] \\ + \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] - \log(\eta) + C(\beta,\gamma)\Bigr] \\ + \mathcal{K}(\rho, \pi) - \mathcal{K}(\pi_{\exp(-\beta R)}, \pi) \\ \leq \mathcal{K}\bigl[\rho, \pi_{\exp(-\frac{\beta\gamma}{N\sinh(\frac{\gamma}{N})}r)}\bigr] \\ + \frac{\beta}{N\sinh(\frac{\gamma}{N})}\Bigl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] + C(\beta,\gamma) - \log(\eta)\Bigr\}.$$

In other words,

Theorem 1.47. *For any positive real constants β and γ such that $\beta < N\sinh(\frac{\gamma}{N})$, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,*

$$\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] \leq \frac{\mathcal{K}\bigl[\rho, \pi_{\exp[-\beta\frac{\gamma}{N}\sinh(\frac{\gamma}{N})^{-1}r]}\bigr]}{1 - \dfrac{\beta}{N\sinh(\frac{\gamma}{N})}} + \frac{C(\beta,\gamma) - \log(\eta)}{\dfrac{N\sinh(\frac{\gamma}{N})}{\beta} - 1}.$$

Choosing in equation (1.29) on page 70 $\alpha = \dfrac{N \log\bigl[\cosh(\frac{\gamma}{N})\bigr]}{1 - \frac{\beta}{N \sinh(\frac{\gamma}{N})}}$ and $\beta = \lambda \frac{N}{\gamma} \sinh(\frac{\gamma}{N})$, so that $\alpha = \dfrac{N \log\bigl[\cosh(\frac{\gamma}{N})\bigr]}{1 - \frac{\lambda}{\gamma}}$, we obtain with $\mathbb{P}$ probability at least $1-\eta$,

$$\begin{aligned}
\log\Bigl\{ \pi_{\exp(-\lambda r)} \Bigl[\exp\bigl\{ N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] \pi_{\exp(-\lambda r)}(m') \bigr\} \Bigr] \Bigr\} \\
\leq \tfrac{2\lambda}{\gamma} \bigl[C(\beta, \gamma) + \log(\tfrac{2}{\eta}) \bigr] \\
+ \Bigl(1 - \tfrac{\lambda}{\gamma}\Bigr) \biggl[\log\Bigl\{ \pi_{\exp(-\beta R)} \otimes \pi_{\exp(-\beta R)} \bigl[\exp(\alpha \Phi_{-\frac{\alpha}{N}} \circ M') \bigr] \Bigr\} \\
+ \log(\tfrac{2}{\eta}) \biggr].
\end{aligned}$$

This proves

Proposition 1.48. *For any positive real constants $\lambda < \gamma$, with $\mathbb{P}$ probability at least $1-\eta$,*

$$\begin{aligned}
\log\Bigl\{ \pi_{\exp(-\lambda r)} \Bigl[\exp\bigl\{ N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] \pi_{\exp(-\lambda r)}(m') \bigr\} \Bigr] \Bigr\} \\
\leq \frac{2\lambda}{\gamma} \bigl[C(\tfrac{N\lambda}{\gamma} \sinh(\tfrac{\gamma}{N}), \gamma) + \log(\tfrac{2}{\eta}) \bigr] \\
+ \Bigl(1 - \tfrac{\lambda}{\gamma}\Bigr) \log\Biggl\{ \pi^{\otimes 2}_{\exp[-\frac{N\lambda}{\gamma} \sinh(\frac{\gamma}{N}) R]} \Biggl[\\
\exp\Biggl(\frac{N \log[\cosh(\frac{\gamma}{N})]}{1 - \frac{\lambda}{\gamma}} \Phi_{-\frac{\log[\cosh(\frac{\gamma}{N})]}{1 - \frac{\lambda}{\gamma}}} \circ M' \Biggr) \Biggr] \Biggr\} \\
+ \Bigl(1 - \tfrac{\lambda}{\gamma}\Bigr) \log(\tfrac{2}{\eta}).
\end{aligned}$$

We are now ready to analyse the bound $B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma)$ of Theorem 1.39 on page 63.

Theorem 1.49. *For any positive real constants λ_1, λ_2, β_1, β_2, β and γ, such that*

$$\begin{aligned}
\lambda_1 &< \gamma, & \beta_1 &< \tfrac{N\lambda_1}{\gamma} \sinh(\tfrac{\gamma}{N}), \\
\lambda_2 &< \gamma, & \beta_2 &> \tfrac{N\lambda_2}{\gamma} \sinh(\tfrac{\gamma}{N}), \\
& & \beta &< \tfrac{N\lambda_2}{\gamma} \tanh(\tfrac{\gamma}{N}),
\end{aligned}$$

with $\mathbb{P}$ probability $1-\eta$, the bound $B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma)$ of Theorem 1.39 on page 63 satisfies

$$
\begin{aligned}
B(&\pi_{\exp(-\lambda_1 r)}, \beta, \gamma) \\
&\leq (\gamma-\lambda_1)\Biggl\{ \pi_{\exp(-\beta_1 R)} \otimes \pi_{\exp(-\beta_2 R)} \bigl[\Psi_{-\frac{\gamma}{N}}(R', M') \bigr] + \frac{\log(\frac{7}{\eta})}{\gamma} \\
&\qquad\qquad + \frac{C(\beta_1,\gamma)+\log(\frac{7}{\eta})}{\frac{N\lambda_1}{\beta_1}\sinh(\frac{\gamma}{N})-\gamma} + \frac{C(\beta_2,\gamma)+\log(\frac{7}{\eta})}{\gamma - \frac{N\lambda_2}{\beta_2}\sinh(\frac{\gamma}{N})} \Biggr\} \\
&\qquad + \frac{2\lambda_1}{\gamma}\Bigl[C\bigl(\tfrac{N\lambda_1}{\gamma}\sinh(\tfrac{\gamma}{N}),\gamma\bigr) + \log(\tfrac{7}{\eta}) \Bigr] \\
&\quad + \Bigl(1-\tfrac{\lambda_1}{\gamma}\Bigr)\log\Biggl\{ \pi^{\otimes 2}_{\exp[-\frac{N\lambda_1}{\gamma}\sinh(\frac{\gamma}{N})R]}\Biggl[\\
&\qquad\qquad \exp\Biggl(\tfrac{N\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_1}{\gamma}} \Phi_{-\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_1}{\gamma}}} \circ M' \Biggr) \Biggr] \Biggr\} \\
&\qquad + \Bigl(1-\tfrac{\lambda_1}{\gamma}\Bigr)\log(\tfrac{7}{\eta}) - \log\bigl[\nu(\{\beta\})\nu(\{\gamma\})\epsilon\bigr] \\
&\quad + (\gamma-\lambda_1)\tfrac{\beta}{\lambda_2} F^{-1}_{\gamma, \frac{\beta\gamma}{\lambda_2}} \Biggl\{ \frac{2\lambda_2}{\gamma}\Bigl[C\bigl(\tfrac{N\lambda_2}{\gamma}\sinh(\tfrac{\gamma}{N}),\gamma\bigr) + \log\bigl(\tfrac{7}{\eta}\bigr) \Bigr] \\
&\qquad + \Bigl(1-\tfrac{\lambda_2}{\gamma}\Bigr)\log\Biggl\{ \pi^{\otimes 2}_{\exp[-\frac{N\lambda_2}{\gamma}\sinh(\frac{\gamma}{N})R]}\Biggl[\\
&\qquad\qquad \exp\Biggl(\frac{N\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_2}{\gamma}} \Phi_{-\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_2}{\gamma}}} \circ M' \Biggr) \Biggr] \Biggr\} \\
&\qquad\qquad + \Bigl(1-\tfrac{\lambda_2}{\gamma}\Bigr)\log\bigl(\tfrac{7}{\eta}\bigr) - \log\bigl[\nu(\{\beta\})\nu(\{\gamma\})\epsilon\bigr] \Biggr\},
\end{aligned}
$$

where the function $C(\beta,\gamma)$ is defined by equation (1.28) *on page 67.*

1.5.4. Adaptation to parametric and margin assumptions. To help understand the previous theorem, it may be useful to give linear upper-bounds to the factors appearing in the right-hand side of the previous inequality. Introducing $\widetilde{\theta}$ such that $R(\widetilde{\theta}) = \inf_\Theta R$ (assuming that such a parameter exists) and remembering that

$$
\begin{aligned}
\Psi_{-a}(p,m) &\leq a^{-1}\sinh(a)p + 2a^{-1}\sinh(\tfrac{a}{2})^2 m, && a\in\mathbb{R}_+, \\
\Phi_{-a}(p) &\leq a^{-1}\bigl[\exp(a)-1\bigr]p, && a\in\mathbb{R}_+, \\
\Psi_{a}(p,m) &\geq a^{-1}\sinh(a)p - 2a^{-1}\sinh(\tfrac{a}{2})^2 m, && a\in\mathbb{R}_+,
\end{aligned}
$$

$$
\begin{aligned}
&M'(\theta_1,\theta_2) \leq M'(\theta_1,\widetilde{\theta}) + M'(\theta_2,\widetilde{\theta}), && \theta_1,\theta_2\in\Theta,\\
&M'(\theta_1,\widetilde{\theta}) \leq xR'(\theta_1,\widetilde{\theta}) + \varphi(x), && x\in\mathbb{R}_+, \theta_1\in\Theta,
\end{aligned}
$$

(the last inequality being rather a consequence of the definition of φ than a property of M'), we easily see that

$$
\begin{aligned}
&\pi_{\exp(-\beta_1 R)}\otimes\pi_{\exp(-\beta_2 R)}\bigl[\Psi_{-\frac{\gamma}{N}}(R',M')\bigr]\\
&\quad\leq \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\bigl[\pi_{\exp(-\beta_1 R)}(R)-\pi_{\exp(-\beta_2 R)}(R)\bigr]\\
&\qquad\qquad\qquad + \tfrac{2N}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\pi_{\exp(-\beta_1 R)}\otimes\pi_{\exp(-\beta_2 R)}(M')\\
&\quad\leq \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\bigl[\pi_{\exp(-\beta_1 R)}(R)-\pi_{\exp(-\beta_2 R)}(R)\bigr]\\
&\qquad + \frac{2xN}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\Bigl\{\pi_{\exp(-\beta_1 R)}\bigl[R'(\cdot,\widetilde{\theta})\bigr]+\pi_{\exp(-\beta_2 R)}\bigl[R'(\cdot,\widetilde{\theta})\bigr]\Bigr\}\\
&\qquad\qquad\qquad\qquad + \frac{4N}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\varphi(x).
\end{aligned}
$$

$$
\begin{aligned}
C(\beta,\gamma) &\leq \log\Bigl\{\pi_{\exp(-\beta R)}\Bigl\{\exp\Bigl[2N\sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2\pi_{\exp(-\beta R)}(M')\Bigr]\Bigr\}\Bigr\}\\
&\leq \log\Bigl\{\pi_{\exp(-\beta R)}\Bigl\{\exp\Bigl[2N\sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 M'(\cdot,\widetilde{\theta})\Bigr]\Bigr\}\Bigr\}\\
&\qquad\qquad + 2N\sinh(\tfrac{\gamma}{2N})^2\pi_{\exp(-\beta R)}\bigl[M'(\cdot,\widetilde{\theta})\bigr]\\
&\leq \log\Bigl\{\pi_{\exp(-\beta R)}\Bigl\{\exp\Bigl[2xN\sinh(\tfrac{\gamma}{2N})^2 R'(\cdot,\widetilde{\theta})\Bigr]\Bigr\}\Bigr\}\\
&\qquad + 2xN\sinh(\tfrac{\gamma}{2N})^2\pi_{\exp(-\beta R)}\bigl[R'(\cdot,\widetilde{\theta})\bigr] + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x)\\
&= \int_{\beta-2xN\sinh(\frac{\gamma}{2N})^2}^{\beta}\pi_{\exp(-\alpha R)}\bigl[R'(\cdot,\widetilde{\theta})\bigr]d\alpha\\
&\qquad + 2xN\sinh(\tfrac{\gamma}{2N})^2\pi_{\exp(-\beta R)}\bigl[R'(\cdot,\widetilde{\theta})\bigr] + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x)\\
&\leq 4xN\sinh(\tfrac{\gamma}{2N})^2\pi_{\exp[-(\beta-2xN\sinh(\frac{\gamma}{2N})^2)R]}\bigl[R'(\cdot,\widetilde{\theta})\bigr]\\
&\qquad\qquad\qquad\qquad + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x).
\end{aligned}
$$

$$
\begin{aligned}
&\log\Bigl\{\pi_{\exp(-\beta R)}^{\otimes 2}\Bigl[\exp\Bigl(N\alpha\Phi_{-\alpha}\circ M'\Bigr)\Bigr]\Bigr\}\\
&\qquad\leq 2\log\Bigl\{\pi_{\exp(-\beta R)}\Bigl[\exp\Bigl(N\bigl[\exp(\alpha)-1\bigr]M'(\cdot,\widetilde{\theta})\Bigr)\Bigr]\Bigr\}
\end{aligned}
$$

$$\leq 2xN\bigl[\exp(\alpha)-1\bigr]\pi_{\exp[-(\beta-xN[\exp(\alpha)-1])R]}\bigl[R'(\cdot,\widetilde{\theta})\bigr] \\ + 2xN\bigl[\exp(\alpha)-1\bigr]\varphi(x).$$

Let us push further the investigation under the parametric assumption that for some positive real constant d

$$\lim_{\beta\to+\infty}\beta\pi_{\exp(-\beta R)}\bigl[R'(\cdot,\widetilde{\theta})\bigr]=d, \tag{1.30}$$

This assumption will for instance hold true with $d=\frac{n}{2}$ when $R:\Theta\to(0,1)$ is a smooth function defined on a compact subset Θ of $\mathbb{R}^n$ that reaches its minimum value on a finite number of non degenerate (i.e. with a positive definite Hessian) interior points of Θ, and π is absolutely continuous with respect to the Lebesgue measure on Θ and has a smooth density.

In case of assumption (1.30), if we restrict to sufficiently large values of the constants β, β_1, β_2, λ_1, λ_2 and γ (the smaller of which being as a rule β as we will see), we can use the fact that for some (small) positive constant δ, and some (large) positive constant A,

$$\frac{d}{\alpha}(1-\delta)\leq\pi_{\exp(-\alpha R)}\bigl[R'(\cdot,\widetilde{\theta})\bigr]\leq\frac{d}{\alpha}(1+\delta),\qquad \alpha\geq A. \tag{1.31}$$

Under this assumption,

$$\pi_{\exp(-\beta_1 R)}\otimes\pi_{\exp(-\beta_2 R)}\bigl[\Psi_{-\frac{\gamma}{N}}(R',M')\bigr] \\ \leq\tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\bigl[\tfrac{d}{\beta_1}(1+\delta)-\tfrac{d}{\beta_2}(1-\delta)\bigr] \\ +\tfrac{2xN}{\gamma}\sinh(\tfrac{\gamma}{2N})^2(1+\delta)\bigl[\tfrac{d}{\beta_1}+\tfrac{d}{\beta_2}\bigr]+\tfrac{4N}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\varphi(x).$$

$$C(\beta,\gamma)\leq d(1+\delta)\log\Bigl(\tfrac{\beta}{\beta-2xN\sinh(\frac{\gamma}{2N})^2}\Bigr) \\ +2xN\sinh(\tfrac{\gamma}{2N})^2\tfrac{(1+\delta)d}{\beta}+4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x).$$

$$\log\Bigl\{\pi^{\otimes 2}_{\exp(-\beta R)}\Bigl[\exp\Bigl(N\alpha\Phi_{-\alpha}\circ M'\Bigr)\Bigr]\Bigr\} \\ \leq 2xN\bigl[\exp(\alpha)-1\bigr]\frac{d(1+\delta)}{\beta-xN[\exp(\alpha)-1]}+2N\bigl[\exp(\alpha)-1\bigr]\varphi(x).$$

Thus with $\mathbb{P}$ probability at least $1-\eta$,

$$B(\pi_{\exp(-\lambda_1 r)},\beta,\gamma)\leq-(\gamma-\lambda_1)\tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\tfrac{d}{\beta_2}(1-\delta) \\ +(\gamma-\lambda_1)\Bigl\{\tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\tfrac{(1+\delta)d}{\beta_1} \\ +\tfrac{2xN}{\gamma}\sinh(\tfrac{\gamma}{2N})^2(1+\delta)\bigl[\tfrac{d}{\beta_1}+\tfrac{d}{\beta_2}\bigr]+\tfrac{4N}{\gamma}\sinh(\tfrac{\gamma}{2N})^2\varphi(x)+\frac{\log(\frac{7}{\eta})}{\gamma}$$

$$
\begin{aligned}
& \quad + \frac{4xN\sinh(\frac{\gamma}{2N})^2 \frac{(1+\delta)d}{\beta_1 - 2xN\sinh(\frac{\gamma}{2N})^2} + 4N\sinh(\frac{\gamma}{2N})^2\varphi(x) + \log(\frac{7}{\eta})}{\frac{N\lambda_1}{\beta_1}\sinh(\frac{\gamma}{N}) - \gamma} \\
& \quad\quad + \frac{4xN\sinh(\frac{\gamma}{2N})^2 \frac{(1+\delta)d}{\beta_2 - 2xN\sinh(\frac{\gamma}{2N})^2} + 4N\sinh(\frac{\gamma}{2N})^2\varphi(x) + \log(\frac{7}{\eta})}{\gamma - \frac{N\lambda_2}{\beta_2}\sinh(\frac{\gamma}{N})} \Biggr\} \\
& + \frac{2\lambda_1}{\gamma}\Biggl\{ 4xN\sinh(\tfrac{\gamma}{2N})^2 \frac{(1+\delta)d}{\frac{N\lambda_1}{\gamma}\sinh(\frac{\gamma}{N}) - 2xN\sinh(\frac{\gamma}{2N})^2} \\
& \qquad\qquad\qquad\qquad + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x) + \log(\tfrac{7}{\eta})\Biggr\} \\
& + \Bigl(1 - \frac{\lambda_1}{\gamma}\Bigr)\Biggl\{ 2d(1+\delta)\left(\frac{\lambda_1 \sinh\bigl(\frac{\gamma}{N}\bigr)}{x\gamma\Bigl[\exp\Bigl(\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_1}{\gamma}}\Bigr) - 1\Bigr]} - 1\right)^{-1} \\
& \qquad\qquad\qquad + 2N\Bigl[\exp\Bigl(\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_1}{\gamma}}\Bigr) - 1\Bigr]\varphi(x)\Biggr\} \\
& \qquad\qquad + \Bigl(1 - \tfrac{\lambda_1}{\gamma}\Bigr)\log(\tfrac{7}{\eta}) - \log\bigl[\nu(\{\beta\})\nu(\{\gamma\})\epsilon\bigr] \\
& + \frac{1 - \frac{\lambda_1}{\gamma}}{\frac{N\lambda_2}{\beta\gamma}\tanh(\frac{\gamma}{N}) - 1}\Biggl\{ \frac{2\lambda_2}{\gamma}\Biggl\{ 4xN\sinh(\tfrac{\gamma}{2N})^2 \frac{(1+\delta)d}{\frac{N\lambda_2}{\gamma}\sinh(\frac{\gamma}{N}) - 2xN\sinh(\frac{\gamma}{2N})^2} \\
& \qquad\qquad\qquad\qquad + 4N\sinh(\tfrac{\gamma}{2N})^2\varphi(x) + \log(\tfrac{7}{\eta})\Biggr\} \\
& + \Bigl(1 - \frac{\lambda_2}{\gamma}\Bigr)\Biggl[2d(1+\delta)\left(\frac{\lambda_2 \sinh\bigl(\frac{\gamma}{N}\bigr)}{x\gamma\Bigl[\exp\Bigl(\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_2}{\gamma}}\Bigr) - 1\Bigr]} - 1\right)^{-1} \\
& \qquad\qquad + 2N\Bigl[\exp\Bigl(\frac{\log[\cosh(\frac{\gamma}{N})]}{1-\frac{\lambda_2}{\gamma}}\Bigr) - 1\Bigr]\varphi(x)\Biggr] \\
& \qquad\qquad\qquad + \Bigl(1 - \tfrac{\lambda_2}{\gamma}\Bigr)\log(\tfrac{7}{\eta}) - \log\bigl[\nu(\beta)\nu(\gamma)\epsilon\bigr]\Biggr\}.
\end{aligned}
$$

Now let us choose for simplicity $\beta_2 = 2\lambda_2 = 4\beta$, $\beta_1 = \lambda_1/2 = \gamma/4$, and let us introduce the notations

$$
\begin{aligned}
C_1 &= \frac{N}{\gamma}\sinh(\frac{\gamma}{N}), \\
C_2 &= \frac{N}{\gamma}\tanh(\frac{\gamma}{N}),
\end{aligned}
$$

$$C_3 = \frac{N^2}{\gamma^2}\bigl[\exp(\frac{\gamma^2}{N^2}) - 1\bigr]$$
$$\text{and} \quad C_4 = \frac{2N^2(1-\frac{2\beta}{\gamma})}{\gamma^2}\Bigl[\exp\Bigl(\frac{\gamma^2}{2N^2(1-\frac{2\beta}{\gamma})}\Bigr) - 1\Bigr],$$

to obtain

$$\begin{aligned}
B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma) \leq {}& -\frac{C_1\gamma}{8\beta}(1-\delta)d \\
&+ \frac{C_1\gamma}{2}\Bigl\{\tfrac{4(1+\delta)d}{\gamma} + x\tfrac{\gamma}{2N}(1+\delta)\bigl[\tfrac{4d}{\gamma} + \tfrac{d}{4\beta}\bigr] + \tfrac{\gamma}{N}\varphi(x)\Bigr\} + \tfrac{1}{2}\log\bigl(\tfrac{7}{\eta}\bigr) \\
&+ \frac{1}{2C_1 - 1}\Bigl[(1+\delta)d\Bigl(\tfrac{N}{2xC_1\gamma} - 1\Bigr)^{-1} + C_1\frac{\gamma^2}{2N}\varphi(x) + \tfrac{1}{2}\log(\tfrac{7}{\eta})\Bigr] \\
&+ \frac{1}{2 - C_1}\Bigl[2(1+\delta)d\Bigl(\tfrac{8N\beta}{xC_1\gamma^2} - 1\Bigr)^{-1} + C_1\frac{\gamma^2}{N}\varphi(x) + \log(\tfrac{7}{\eta})\Bigr] \\
&\qquad + \frac{2x\gamma(1+\delta)d}{N - x\gamma} + C_1\tfrac{\gamma^2}{N}\varphi(x) + \log(\tfrac{7}{\eta}) \\
&+ d(1+\delta)\frac{x\gamma}{N}\Bigl(\frac{C_1}{2C_3} - \frac{x\gamma}{N}\Bigr)^{-1} + \frac{\gamma^2}{N}C_3\varphi(x) + \frac{\log(\frac{7}{\eta})}{2} - \log\bigl[\nu(\beta)\nu(\gamma)\epsilon\bigr] \\
&+ \Bigl(4C_2 - 2\Bigr)^{-1}\Biggl\{\frac{4\beta}{\gamma}\Bigl\{x\frac{\gamma^2}{N}C_1(1+\delta)d\Bigl(2\beta C_1 - xC_1\frac{\gamma^2}{2N}\Bigr)^{-1} \\
&\qquad + \tfrac{\gamma^2}{N}\varphi(x) + \log(\tfrac{7}{\eta})\Bigr\} \\
&+ \Bigl(1 - \frac{2\beta}{\gamma}\Bigr)\Bigl\{2d(1+\delta)\frac{x\gamma}{N}\Bigl[\frac{4\beta C_1}{\gamma C_4}\Bigl(1 - \frac{2\beta}{\gamma}\Bigr) - \frac{x\gamma}{N}\Bigr]^{-1} \\
&\qquad + \frac{\gamma^2}{N(1-\frac{2\beta}{\gamma})}C_4\varphi(x)\Bigr\} \\
&\qquad + \Bigl(1 - \tfrac{2\beta}{\gamma}\Bigr)\log(\tfrac{7}{\eta}) - \log\bigl[\nu(\beta)\nu(\gamma)\epsilon\bigr]\Biggr\}.
\end{aligned}$$

This simplifies to

$$\begin{aligned}
B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma) \leq {}& -\frac{C_1}{8}(1-\delta)d\frac{\gamma}{\beta} \\
&+ 2C_1(1+\delta)d + \log(\tfrac{7}{\eta})\Bigl[2 + \tfrac{3C_1}{(4C_1-2)(2-C_1)} + \frac{1 + \frac{2\beta}{\gamma}}{4C_2 - 2}\Bigr]
\end{aligned}$$

$$
\begin{aligned}
& \qquad\qquad\qquad\qquad\qquad\qquad - \Bigl(1+\tfrac{1}{4C_2-2}\Bigr)\log\bigl[\nu(\beta)\nu(\gamma)\epsilon\bigr] \\
& + \frac{(1+\delta)dx\gamma}{N}\biggl\{C_1 + \tfrac{1}{2C_1-1}\Bigl(\tfrac{1}{2C_1}-\tfrac{\gamma x}{N}\Bigr)^{-1} \\
& \qquad\qquad\qquad + 2\Bigl(1-\tfrac{\gamma x}{N}\Bigr)^{-1} + \Bigl(\tfrac{C_1}{2C_3}-\tfrac{\gamma x}{N}\Bigr)^{-1} + \tfrac{4C_1\beta}{\gamma(4C_2-2)}\biggr\} \\
& + \frac{(1+\delta)dx\gamma^2}{N\beta}\biggl\{\tfrac{C_1}{16} + \tfrac{2}{2-C_1}\Bigl(\tfrac{8}{C_1}-\tfrac{x\gamma^2}{N\beta}\Bigr)^{-1} \\
& \qquad\qquad\qquad + \Bigl(1-\tfrac{2\beta}{\gamma}\Bigr)\tfrac{1}{2C_2-1}\Bigl[\tfrac{4C_1}{C_4}\Bigl(1-\tfrac{2\beta}{\gamma}\Bigr)-\tfrac{\gamma^2 x}{\beta N}\Bigr]^{-1}\biggr\} \\
& \qquad + \frac{\gamma^2}{N}\varphi(x)\biggl\{\tfrac{3C_1}{2} + \tfrac{C_1}{4C_1-2} + \tfrac{C_1}{2-C_1} + C_3 + \tfrac{4\beta}{\gamma(4C_2-2)} + \tfrac{C_4}{4C_2-2}\biggr\}.
\end{aligned}
$$

This shows that there exist universal positive real constants A_1, A_2, B_1, B_2, B_3, and B_4 such that as soon as $\frac{\gamma \max\{x,1\}}{N} \leq A_1 \frac{\beta}{\gamma} \leq A_2$,

$$
\begin{aligned}
B(\pi_{\exp(-\lambda_1 r)}, \beta, \gamma) \leq -B_1(1-\delta)d\frac{\gamma}{\beta} + B_2(1+\delta)d \\
- B_3 \log\bigl[\nu(\beta)\nu(\gamma)\epsilon\,\eta\bigr] + B_4\frac{\gamma^2}{N}\varphi(x).
\end{aligned}
$$

Thus $\pi_{\exp(-\lambda_1 r)}(R) \leq \pi_{\exp(-\beta R)}(R) \leq \inf_\Theta R + \frac{(1+\delta)d}{\beta}$ as soon as moreover

$$
\frac{\beta}{\gamma} \leq \frac{B_1}{B_2\frac{(1+\delta)}{(1-\delta)} + \frac{B_4\frac{\gamma^2}{N}\varphi(x) - B_3\log[\nu(\beta)\nu(\gamma)\epsilon\eta]}{(1-\delta)d}}.
$$

Choosing some real ratio $\alpha > 1$, we can now make the above result uniform for any

$$
\beta, \gamma \in \Lambda_\alpha \overset{\text{def}}{=} \Bigl\{\alpha^k; k \in \mathbb{N}, 0 \leq k < \tfrac{\log(N)}{\log(\alpha)}\Bigr\}, \tag{1.32}
$$

by substituting $\nu(\beta)$ and $\nu(\gamma)$ with $\frac{\log(\alpha)}{\log(\alpha N)}$ and $-\log(\eta)$ with $-\log(\eta) + 2\log\Bigl[\frac{\log(\alpha N)}{\log(\alpha)}\Bigr]$.

Taking moreover for simplicity $\eta = \epsilon$, let us summarize the type of result we got by

Theorem 1.50. *There exist positive real universal constants A, B_1, B_2, B_3 and B_4 such that for any positive real constants $\alpha > 1$, d and δ, for any prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\beta, \gamma \in \Lambda_\alpha$ (where Λ_α is defined by equation (1.32) above) such that*

$$
\sup_{\beta' \in \mathbb{R}, \beta' \geq \beta} \biggl| \frac{\beta'}{d}\bigl[\pi_{\exp(-\beta' R)}(R) - \inf_\Theta R\bigr] - 1 \biggr| \leq \delta
$$

and such that also for some positive real parameter x

$$\frac{\gamma \max\{x, 1\}}{N} \leq \frac{A\beta}{\gamma} \text{ and } \frac{\beta}{\gamma} \leq \frac{B_1}{B_2 \frac{(1+\delta)}{(1-\delta)} + \frac{B_4 \frac{\gamma^2}{N}\varphi(x) - 2B_3 \log(\epsilon) + 4B_3 \log\left[\frac{\log(N)}{\log(\alpha)}\right]}{(1-\delta)d}},$$

the bound $B(\pi_{\exp(-\frac{\gamma}{2}r)}, \beta, \gamma)$ *given by Theorem 1.39 on page 63 in the case where we have chosen* ν *to be the uniform probability measure on* Λ_α*, satisfies* $B(\pi_{\exp(-\frac{\gamma}{2}r)}, \beta, \gamma) \leq 0$*, proving that* $\widehat{\beta}(\pi_{\exp(-\frac{\gamma}{2}r)}) \geq \beta$ *and therefore that*

$$\pi_{\exp(-\gamma\frac{r}{2})}(R) \leq \pi_{\exp(-\beta R)}(R) \leq \inf_\Theta R + \frac{(1+\delta)d}{\beta}.$$

What is important in this result is that we do not only bound $\pi_{\exp(-\frac{\gamma}{2}r)}(R)$, but also $B(\pi_{\exp(-\frac{\gamma}{2}r)}, \beta, \gamma)$, and that we do it uniformly on a grid of values of β and γ, showing that we can indeed set the constants β and γ adaptively using the empirical bound $B(\pi_{\exp(-\frac{\gamma}{2}r)}, \beta, \gamma)$.

Let us see what we get under the margin assumption (1.21) (see page 47). When $\kappa = 1$, $\varphi(c^{-1}) \leq 0$, leading to

COROLLARY 1.51. *Assuming that the margin assumption 1.21 (on page 47) is satisfied for* $\kappa = 1$*, that* $R : \Theta \to (0, 1)$ *is independent of* N *(which is the case for instance when* $\mathbb{P} = P^{\otimes N}$*), and is such that*

$$\lim_{\beta' \to +\infty} \beta' \bigl[\pi_{\exp(-\beta' R)}(R) - \inf_\Theta R\bigr] = d,$$

there are universal positive real constants B_5 *and* B_6 *and* $N_1 \in \mathbb{N}$ *such that for any* $N \geq N_1$*, with* $\mathbb{P}$ *probability at least* $1 - \epsilon$

$$\pi_{\exp(-\widehat{\gamma}\frac{r}{2})}(R) \leq \inf_\Theta R + \frac{B_5 d}{cN}\left[1 + \frac{B_6}{d}\log\left(\frac{\log(N)}{\epsilon}\right)\right]^2,$$

where $\widehat{\gamma} \in \arg\max_{\gamma \in \Lambda_2} \max\bigl\{\beta \in \Lambda_2; B(\pi_{\exp(-\gamma\frac{r}{2})}, \beta, \gamma) \leq 0\bigr\}$ *(where* Λ_2 *is defined by equation* (1.32) *on page 77).*

When $\kappa > 1$, $\varphi(x) \leq (1 - \kappa^{-1})\bigl(\kappa c x\bigr)^{-\frac{1}{\kappa - 1}}$, and we can choose γ and x such that $\frac{\gamma^2}{N}\varphi(x) \simeq d$ to prove

COROLLARY 1.52. *Assuming that the margin assumption* (1.21) *is satisfied for some exponent* $\kappa > 1$*, that* $R : \Theta \to (0, 1)$ *is independent of* N *(which is for instance the case when* $\mathbb{P} = P^{\otimes N}$*), and is such that*

$$\lim_{\beta' \to +\infty} \beta' \bigl[\pi_{\exp(-\beta' R)}(R) - \inf_\Theta R\bigr] = d,$$

there are universal positive constants B_7 *and* B_8 *and* $N_1 \in \mathbb{N}$ *such that for any* $N \geq N_1$*, with* $\mathbb{P}$ *probability at least* $1 - \epsilon$*,*

$$\pi_{\exp(-\widehat{\gamma}\frac{r}{2})}(R) \leq \inf_{\Theta} R + B_7 c^{-\frac{1}{2\kappa-1}} \left[1 + \frac{B_8}{d} \log \left(\frac{\log(N)}{\epsilon} \right) \right]^{\frac{2\kappa}{2\kappa-1}} \left(\frac{d}{N} \right)^{\frac{\kappa}{2\kappa-1}},$$

where $\widehat{\gamma} \in \arg\max_{\gamma \in \Lambda_2} \max\bigl\{\beta \in \Lambda_2; B(\pi_{\exp(-\gamma\frac{r}{2})}, \beta, \gamma) \leq 0\bigr\}$ *(Λ_2 being defined by equation* (1.32) *on page 77).*

We find the same rate of convergence as in Corollary 1.30 on page 48, but this time, we were able to provide an empirical posterior distribution $\pi_{\exp(-\widehat{\gamma}\frac{r}{2})}$ which achieves this rate adaptively in all the parameters (meaning in particular that we do not need to know d, c or κ). Moreover, as already mentioned, the power of N in this rate of convergence is known to be unimprovable in the worst case (see [28, 34, 35], and more specifically in [3] — downloadable from its author's web page,— Theorem 3.3 on page 132).

1.5.5. Estimating the divergence of a posterior with respect to a Gibbs prior. Another interesting question is to estimate $\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr]$ using relative deviation inequalities. We follow here an idea to be found first in Audibert [3, page 93]. Indeed, combining equation (1.24) with equation (1.25) on page 61, we see that for any positive real parameters β and λ, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}
\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] \leq \frac{\beta}{N\lambda} \biggl\{ \frac{N}{2} \log\left(\frac{1+\lambda}{1-\lambda}\right) \bigl[\rho(r) - \pi_{\exp(-\beta R)}(r)\bigr] \\
- \frac{N}{2} \log(1-\lambda^2) \rho \otimes \pi_{\exp(-\beta R)}(m') \\
+ \mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] - \log(\epsilon) \biggr\} + \mathcal{K}(\rho, \pi) - \mathcal{K}\bigl[\pi_{\exp(-\beta R)}, \pi\bigr] \\
\leq \mathcal{K}\bigl[\rho, \pi_{\exp[-\frac{\beta}{2\lambda}\log(\frac{1+\lambda}{1-\lambda})r]}\bigr] + \frac{\beta}{N\lambda}\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] - \frac{\beta}{N\lambda}\log(\epsilon) \\
+ \log\biggl[\pi_{\exp[-\frac{\beta}{2\lambda}\log(\frac{1+\lambda}{1-\lambda})r]}\Bigl\{\exp\Bigl[-\frac{\beta}{2\lambda}\log(1-\lambda^2)\rho(m')\Bigr]\Bigr\}\biggr].
\end{aligned}$$

Thus, putting $\gamma = \frac{N}{2}\log(\frac{1+\lambda}{1-\lambda})$, we obtain

THEOREM 1.53. *For any positive real constants* β *and* γ *such that* $\beta < N\tanh(\frac{\gamma}{N})$*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any posterior distribution* $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$*,*

$$\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta R)}\bigr] \leq \left(1 - \frac{\beta}{N} \tanh\left(\frac{\gamma}{N}\right)^{-1}\right)^{-1} \\ \times \biggl\{ \mathcal{K}\bigl[\rho, \pi_{\exp[-\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1} r]}\bigr] - \frac{\beta}{N\tanh(\frac{\gamma}{N})}\log(\epsilon) \\ + \log\Bigl\{ \pi_{\exp[-\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1} r]}\Bigl[\exp\bigl\{\beta\tanh(\tfrac{\gamma}{N})^{-1}\log[\cosh(\tfrac{\gamma}{N})]\rho(m')\bigr\}\Bigr]\Bigr\}\biggr\}.$$

This theorem provides another way of measuring overfitting, since it gives an upper bound for $\mathcal{K}\bigl[\pi_{\exp[-\frac{\beta\gamma}{N}\tanh(\frac{\gamma}{N})^{-1} r]}, \pi_{\exp(-\beta R)}\bigr]$. It may be used in combination with Theorem 1.10 on page 20 as an alternative to Theorem 1.18 on page 30. It will also be used in the next section.

An alternative parametrization of the same result providing a simpler right-hand side is also useful:

COROLLARY 1.54 *For any positive real constants β and γ such that $\beta < \gamma$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\mathcal{K}\bigl[\rho, \pi_{\exp[-N\frac{\beta}{\gamma}\tanh(\frac{\gamma}{N})R]}\bigr] \leq \left(1 - \frac{\beta}{\gamma}\right)^{-1}\biggl\{\mathcal{K}\bigl[\rho, \pi_{\exp(-\beta r)}\bigr] - \frac{\beta}{\gamma}\log(\epsilon) \\ + \log\Bigl\{\pi_{\exp(-\beta r)}\Bigl[\exp\bigl\{N\tfrac{\beta}{\gamma}\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\rho(m')\bigr\}\Bigr]\Bigr\}\biggr\}.$$

1.5.6. Comparing two posterior distributions. Estimating the effective temperature of an estimator provides an efficient way to tune parameters in a model with a parametric behaviour. On the other hand, it will not be fitted to choose between different models, especially in the case when they are nested (because as we already saw in the case when Θ is a union of nested models, the prior distribution $\pi_{\exp(-\beta R)}$ is not providing an efficient localization of the parameter in this case, in the sens that $\pi_{\exp(-\beta R)}(R)$ is not going down to $\inf_\Theta R$ at the desired rate when β goes to $+\infty$, requiring to resort to partial localization).

Once some estimator (in the form of a posterior distribution) has been chosen in each submodel, these estimators can be compared between themselves with the help of the relative bounds that we will establish in this section.

From equation (1.23) (slightly modified by replacing $\pi \otimes \pi$ with $\pi^1 \otimes \pi^2$), we obtain easily

Theorem 1.55. *For any positive real constant λ, for any prior distributions $\pi^1, \pi^2 \in \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distributions ρ_1 and $\rho_2 : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\begin{aligned} -N \log\Bigl\{1 - \tanh\bigl(\tfrac{\lambda}{N}\bigr)\Bigl[\rho_2(R) - \rho_1(R)\Bigr]\Bigr\} \leq \lambda\bigl[\rho_2(r) - \rho_1(r)\bigr] \\ + N \log\bigl[\cosh\bigl(\tfrac{\lambda}{N}\bigr)\bigr] \rho_1 \otimes \rho_2(m') \\ + \mathcal{K}\bigl(\rho_1, \pi^1\bigr) + \mathcal{K}\bigl(\rho_2, \pi^2\bigr) - \log(\epsilon). \end{aligned}$$

There enters into the game the entropy bound of the previous section, providing a localized version of Theorem 1.55. We will use the notation

$$\Xi_a(q) = \tanh(a)^{-1}\bigl[1 - \exp(-aq)\bigr] \leq \frac{a}{\tanh(a)} q, \qquad a, q \in \mathbb{R}.$$

Theorem 1.56. *For any sequence of prior distributions $(\pi^i)_{i \in \mathbb{N}} \in \mathcal{M}_+^1(\Theta)^{\mathbb{N}}$, any probability distribution μ on $\mathbb{N}$, any atomic probability distribution ν on $\mathbb{R}_+$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any posterior distributions $\rho_1, \rho_2 : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\rho_2(R) - \rho_1(R) \leq B(\rho_1, \rho_2), \text{ where}$$

$$\begin{aligned} B(\rho_1, \rho_2) = \inf_{\lambda, \beta_1 < \gamma_1, \beta_2 < \gamma_2 \in \mathbb{R}_+, i, j \in \mathbb{N}} \Xi_{\frac{\lambda}{N}} \Biggl\{ \bigl[\rho_2(r) - \rho_1(r)\bigr] \\ + \tfrac{N}{\lambda} \log\bigl[\cosh(\tfrac{\lambda}{N})\bigr] \rho_1 \otimes \rho_2(m') \\ + \frac{1}{\lambda\Bigl(1 - \frac{\beta_1}{\gamma_1}\Bigr)} \Biggl\{ \mathcal{K}\bigl[\rho_1, \pi^i_{\exp(-\beta_1 r)}\bigr] \\ + \log\Bigl\{\pi^i_{\exp(-\beta_1 r)}\Bigl[\exp\bigl\{\beta_1 \tfrac{N}{\gamma_1} \log\bigl[\cosh(\tfrac{\gamma_1}{N})\bigr] \rho_1(m')\bigr\}\Bigr]\Bigr\}\Biggr\} \\ + \frac{1}{\lambda\Bigl(1 - \frac{\beta_2}{\gamma_2}\Bigr)} \Biggl\{ \mathcal{K}\bigl[\rho_2, \pi^j_{\exp(-\beta_2 r)}\bigr] \\ + \log\Bigl\{\pi^j_{\exp(-\beta_2 r)}\Bigl[\exp\bigl\{\beta_2 \tfrac{N}{\gamma_2} \log\bigl[\cosh(\tfrac{\gamma_2}{N})\bigr] \rho_2(m')\bigr\}\Bigr]\Bigr\}\Biggr\} \\ - \Bigl[\bigl(\tfrac{\gamma_1}{\beta_1} - 1\bigr)^{-1} + \bigl(\tfrac{\gamma_2}{\beta_2} - 1\bigr)^{-1} + 1\Bigr] \\ \times \frac{\log\bigl[3^{-1}\nu(\beta_1)\nu(\beta_2)\nu(\gamma_1)\nu(\gamma_2)\nu(\lambda)\mu(i)\mu(j)\epsilon\bigr]}{\lambda} \Biggr\}. \end{aligned}$$

The sequence of prior distributions $(\pi^i)_{i\in\mathbb{N}}$ should be understood to be typically supported by subsets of Θ corresponding to parametric submodels, that is submodels for which it is reasonable to expect that

$$\lim_{\beta\to+\infty} \beta\bigl[\pi^i_{\exp(-\beta R)}(R) - \operatorname*{ess\,inf}_{\pi^i} R\bigr]$$

exists and is positive and finite. As there is no reason why the bound $B(\rho_1,\rho_2)$ provided by the previous theorem should be subadditive (in the sense that $B(\rho_1,\rho_3) \leq B(\rho_1,\rho_2) + B(\rho_2,\rho_3)$), it is adequate, at least from a theoretical point of view, to consider some workable subset $\mathcal{P} \subset \mathcal{M}^1_+(\Theta)$ of posterior distributions (for instance the distributions of the form $\pi^i_{\exp(-\beta r)}$, $i \in \mathbb{N}$, $\beta \in \mathbb{R}_+$, it is understood that $\mathcal{P}$ is allowed to be a random subset of $\mathcal{M}^1_+(\Theta)$, as in this suggested example), and to define the subadditive chained bound

$$\widetilde{B}(\rho,\rho') = \inf\Biggl\{\sum_{k=0}^{n-1} B(\rho_k,\rho_{k+1});\, n\in\mathbb{N}^*, (\rho_k)_{k=0}^n \in \mathcal{P}^{n+1},\\ \rho_0=\rho, \rho_n=\rho'\Biggr\}, \quad \rho,\rho'\in\mathcal{P}.$$

Proposition 1.57. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\rho_1,\rho_2 \in \mathcal{P}$, $\rho_2(R) - \rho_1(R) \leq \widetilde{B}(\rho_1,\rho_2)$. Moreover for any posterior distribution $\rho_1 \in \mathcal{P}$, any posterior distribution $\rho_2 \in \mathcal{P}$ such that $\widetilde{B}(\rho_1,\rho_2) = \inf_{\rho_3\in\mathcal{P}} \widetilde{B}(\rho_1,\rho_3)$ is unimprovable with the help of $\widetilde{B}$ in $\mathcal{P}$ in the sense that $\inf_{\rho_3\in\mathcal{P}} \widetilde{B}(\rho_2,\rho_3) \geq 0$.*

Proof. The first assertion is a direct consequence of the previous theorem, therefore only the second assertion requires a proof: for any $\rho_3 \in \mathcal{P}$, we deduce from the optimality of ρ_2 and the subadditivity of $\widetilde{B}$ that $\widetilde{B}(\rho_1,\rho_2) \leq \widetilde{B}(\rho_1,\rho_3) \leq \widetilde{B}(\rho_1,\rho_2) + \widetilde{B}(\rho_2,\rho_3)$. $\square$

This proposition provides a way to improve a posterior distribution $\rho_1 \in \mathcal{P}$ by choosing $\rho_2 \in \arg\min_{\rho\in\mathcal{P}} \widetilde{B}(\rho_1,\rho)$ whenever $\widetilde{B}(\rho_1,\rho_2) < 0$. This improvement process is proved according to Proposition 1.57 to be a one step process: the obtained improved posterior ρ_2 cannot be improved again using the same technique.

Let us give some example of possible starting distribution ρ_1 for this improvement scheme: ρ_1 may be chosen as the best posterior Gibbs distribution according to Proposition 1.40 on page 66. More precisely, we may build from the prior distributions π^i, $i\in\mathbb{N}$, a global prior $\pi = \sum_{i\in\mathbb{N}} \mu(i)\pi^i$. We can

then define the estimator of the inverse effective temperature as in Proposition 1.40 and choose $\rho_1 \in \arg\min_{\rho \in \mathcal{P}} \widehat{\beta}(\rho)$, where $\mathcal{P}$ is as suggested above the set of posterior distributions

$$\mathcal{P} = \Bigl\{ \pi^i_{\exp(-\beta r)} ; \, i \in \mathbb{N}, \beta \in \mathbb{R}_+ \Bigr\}.$$

(This starting point ρ_1 should already be pretty good, at least in an asymptotic perspective, the only gain in the rate of convergence to be expected bearing on spurious $\log(N)$ factors).

For more elaborate uses of relative bounds, we refer to the third section of the second chapter of Audibert [3], where an algorithm is proposed and analyzed, which allows to use relative bounds between two posterior distributions as a stand alone estimation tool.

1.5.7. Two step localization of relative bounds. Let us consider again in this section the case when we want to choose adaptively between a family of parametric models. Let us thus assume that the parameter set is a disjoint union of measurable submodels, so that we can write $\Theta = \sqcup_{m \in M} \Theta_m$, where M is some measurable index set. Let us choose some prior probability distribution on the index set $\mu \in \mathcal{M}^1_+(M)$, and some regular conditional prior distribution on (M, Θ), $\pi : M \to \mathcal{M}^1_+(\Theta)$, such that $\pi(m, \Theta_m) = 1$, $m \in M$. Let us then study some arbitrary posterior distributions $\nu : \Omega \to \mathcal{M}^1_+(M)$ and $\rho : \Omega \times M :\to \mathcal{M}^1_+(\Theta)$, such that $\rho(\omega, m, \Theta_m) = 1$, $\omega \in \Omega$, $m \in M$. We would like to compare $\nu\rho(R)$ with some doubly localized prior distribution $\mu_{\exp[-\frac{\beta}{1+\zeta_2}\pi_{\exp(-\beta R)}(R)]}\bigl[\pi_{\exp(-\beta R)}\bigr](R)$ (where ζ_2 is a positive parameter to be set as needed later on). We will define to ease notations two prior distributions (one being more precisely a conditional distribution) depending on the positive real parameters β and ζ_2, putting

$$\overline{\pi} = \pi_{\exp(-\beta R)} \text{ and } \overline{\mu} = \mu_{\exp[-\frac{\beta}{1+\zeta_2}\overline{\pi}(R)]}. \tag{1.33}$$

Similarly to Theorem 1.26 on page 44 we can write for any positive real constants β and γ

$$\mathbb{P}\biggl\{ (\overline{\mu}\,\overline{\pi}) \otimes (\overline{\mu}\,\overline{\pi}) \Bigl[\exp\Bigl[-N \log\bigl[1 - \tanh(\tfrac{\gamma}{N}) R'\bigr] \\ - \gamma r' - N \log\bigl[\cosh(\tfrac{\gamma}{N})\bigr] m' \Bigr]\Bigr] \biggr\} \leq 1,$$

and deduce, using Lemma 1.3 on page 11

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\nu\in\mathcal{M}_+^1(M)}\sup_{\rho:M\to\mathcal{M}_+^1(\Theta)}\Bigl\{-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})(\nu\rho-\overline{\mu}\,\overline{\pi})(R)\bigr] \\ -\gamma(\nu\rho-\overline{\mu}\,\overline{\pi})(r)-N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr](\nu\rho)\otimes(\overline{\mu}\,\overline{\pi})(m') \\ -\mathcal{K}(\nu,\overline{\mu})-\nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr]\Bigr\}\biggr]\biggr\}\leq 1. \quad (1.34)$$

This will be our starting point in comparing $\nu\rho(R)$ with $\overline{\mu}\,\overline{\pi}(R)$. However, obtaining an empirical bound will require some supplementary efforts. For each $m\in M$, we can write in the same way

$$\mathbb{P}\biggl\{\overline{\pi}\otimes\overline{\pi}\biggl[\exp\Bigl[-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})R'\bigr]-\gamma r'-N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]m'\Bigr]\biggr]\biggr\}\leq 1.$$

Intagrating this inequality with respect to $\overline{\mu}$ and using Fubini's lemma for positive functions, we get

$$\mathbb{P}\biggl\{\overline{\mu}(\overline{\pi}\otimes\overline{\pi})\biggl[\exp\Bigl[-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})R'\bigr]-\gamma r'-N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]m'\Bigr]\biggr]\biggr\}\leq 1.$$

Let us make clear that $\overline{\mu}(\overline{\pi}\otimes\overline{\pi})$ is a probability measure on $M\times\Theta\times\Theta$, whereas $(\overline{\mu}\,\overline{\pi})\otimes(\overline{\mu}\,\overline{\pi})$ considered previously is a probability measure on $(M\times\Theta)\times(M\times\Theta)$. We get as previously

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\nu\in\mathcal{M}_+^1(M)}\sup_{\rho:M\to\mathcal{M}_+^1(\Theta)}\Bigl\{-N\log\bigl[1-\tanh(\tfrac{\gamma}{N})\nu(\rho-\overline{\pi})(R)\bigr] \\ -\gamma\nu(\rho-\overline{\pi})(r)-N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\nu(\rho\otimes\overline{\pi})(m') \\ -\mathcal{K}(\nu,\overline{\mu})-\nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr]\Bigr\}\biggr]\biggr\}\leq 1. \quad (1.35)$$

Let us eventually recall that

$$\mathcal{K}(\nu,\overline{\mu})=\tfrac{\beta}{1+\zeta_2}(\nu-\overline{\mu})\overline{\pi}(R)+\mathcal{K}(\nu,\mu)-\mathcal{K}(\overline{\mu},\mu), \quad (1.36)$$

$$\mathcal{K}(\rho,\overline{\pi})=\beta(\rho-\overline{\pi})(R)+\mathcal{K}(\rho,\pi)-\mathcal{K}(\overline{\pi},\pi). \quad (1.37)$$

From equations (1.34), (1.35) and (1.37) we deduce

Proposition 1.58. *For any positive real constants β, γ and ζ_2, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu:\Omega\to\mathcal{M}_+^1(M)$ and any conditional posterior distribution $\rho:\Omega\times M\to\mathcal{M}_+^1(\Theta)$,*

$$
\begin{aligned}
- N \log\bigl[1 - \tanh(\tfrac{\gamma}{N})(\nu\rho - \overline{\mu}\,\overline{\pi})(R)\bigr] - \beta\nu(\rho - \overline{\pi})(R) \\
\leq \gamma(\nu\rho - \overline{\mu}\,\overline{\pi})(r) + N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr](\nu\rho)\otimes(\overline{\mu}\,\overline{\pi})(m') \\
+ \mathcal{K}(\nu,\overline{\mu}) + \nu\bigl[\mathcal{K}(\rho,\pi)\bigr] - \nu\bigl[\mathcal{K}(\overline{\pi},\pi)\bigr] + \log\bigl(\tfrac{2}{\epsilon}\bigr).
\end{aligned}
$$

and

$$
\begin{aligned}
- N \log\bigl[1 - \tanh(\tfrac{\gamma}{N})\nu(\rho - \overline{\pi})(R)\bigr] \\
\leq \gamma\nu(\rho - \overline{\pi})(r) + N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr]\nu(\rho\otimes\overline{\pi})(m') \\
+ \mathcal{K}(\nu,\overline{\mu}) + \nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr] + \log\bigl(\tfrac{2}{\epsilon}\bigr),
\end{aligned}
$$

where the prior distribution $\overline{\mu}\,\overline{\pi}$ *is defined by equation* (1.33) *on page 83 and depends on* β *and* ζ_2*.*

Let us put for short

$$T = \tanh(\tfrac{\gamma}{N}) \text{ and } C = N\log\bigl[\cosh(\tfrac{\gamma}{N})\bigr].$$

We will use some entropy compensation strategy for which we need a couple of entropy bounds. Let us assume that $\beta < NT$. We have according to Proposition 1.58, with $\mathbb{P}$ probability at least $1-\epsilon$,

$$
\begin{aligned}
\nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr] &= \beta\nu(\rho-\overline{\pi})(R) + \nu\bigl[\mathcal{K}(\rho,\pi) - \mathcal{K}(\overline{\pi},\pi)\bigr] \\
&\leq \frac{\beta}{NT}\biggl[\gamma\nu(\rho-\overline{\pi})(r) + C\nu(\rho\otimes\overline{\pi})(m') \\
&\qquad + \mathcal{K}(\nu,\overline{\mu}) + \nu\bigl[\mathcal{K}(\rho,\overline{\pi})\bigr] + \log(\tfrac{2}{\epsilon})\biggr] \\
&\qquad\qquad + \nu\bigl[\mathcal{K}(\rho,\pi) - \mathcal{K}(\overline{\pi},\pi)\bigr].
\end{aligned}
$$

Similarly

$$
\begin{aligned}
\mathcal{K}(\nu,\overline{\mu}) &= \frac{\beta}{1+\zeta_2}(\nu-\overline{\mu})\overline{\pi}(R) + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu) \\
&\leq \frac{\beta}{(1+\zeta_2)NT}\biggl[\gamma(\nu-\overline{\mu})\overline{\pi}(r) + C(\nu\overline{\pi})\otimes(\overline{\mu}\,\overline{\pi})(m') \\
&\qquad + \mathcal{K}(\nu,\overline{\mu}) + \log(\tfrac{2}{\epsilon})\biggr] + \mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu).
\end{aligned}
$$

Thus, for any positive real constants β, γ and ζ_i, $i = 1, \dots, 5$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distributions $\nu, \nu_3 : \Omega \to \mathcal{M}_+^1(\Theta)$, any posterior conditional distributions $\rho, \rho_1, \rho_2, \rho_4, \rho_5 : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,

$$
\begin{aligned}
& - N \log\bigl[1 - T(\nu\rho - \overline{\mu}\,\overline{\pi})(R)\bigr] - \beta\nu(\rho - \overline{\pi})(R) \\
& \qquad \qquad \leq \gamma(\nu\rho - \overline{\mu}\,\overline{\pi})(r) + C(\nu\rho) \otimes (\overline{\mu}\,\overline{\pi})(m') \\
& \qquad \qquad \qquad + \mathcal{K}(\nu, \overline{\mu}) + \nu\bigl[\mathcal{K}(\rho, \pi) - \mathcal{K}(\overline{\pi}, \pi)\bigr] + \log(\tfrac{2}{\epsilon}), \\
& \zeta_1 \frac{NT}{\beta} \overline{\mu}\bigl[\mathcal{K}(\rho_1, \overline{\pi})\bigr] \leq \zeta_1 \gamma \overline{\mu}(\rho_1 - \overline{\pi})(r) + \zeta_1 C \overline{\mu}(\rho_1 \otimes \overline{\pi})(m') \\
& \qquad + \zeta_1 \overline{\mu}\bigl[\mathcal{K}(\rho_1, \overline{\pi})\bigr] + \zeta_1 \log(\tfrac{2}{\epsilon}) + \zeta_1 \frac{NT}{\beta} \overline{\mu}\bigl[\mathcal{K}(\rho_1, \pi) - \mathcal{K}(\overline{\pi}, \pi)\bigr], \\
& \zeta_2 \frac{NT}{\beta} \nu\bigl[\mathcal{K}(\rho_2, \overline{\pi})\bigr] \leq \zeta_2 \gamma \nu(\rho_2 - \overline{\pi})(r) + \zeta_2 C \nu(\rho_2 \otimes \overline{\pi})(m') \\
& \qquad + \zeta_2 \mathcal{K}(\nu, \overline{\mu}) + \zeta_2 \nu\bigl[\mathcal{K}(\rho_2, \overline{\pi})\bigr] + \zeta_2 \log(\tfrac{2}{\epsilon}) \\
& \qquad \qquad + \zeta_2 \frac{NT}{\beta} \nu\bigl[\mathcal{K}(\rho_2, \pi) - \mathcal{K}(\overline{\pi}, \pi)\bigr], \\
& \zeta_3(1 + \zeta_2) \frac{NT}{\beta} \mathcal{K}(\nu_3, \overline{\mu}) \leq \zeta_3 \gamma (\nu_3 - \overline{\mu}) \overline{\pi}(r) \\
& \quad + \zeta_3 C\bigl[(\nu_3 \overline{\pi}) \otimes (\nu_3 \rho_1) + (\nu_3 \rho_1) \otimes (\overline{\mu}\,\overline{\pi})\bigr](m') + \zeta_3 \mathcal{K}(\nu_3, \overline{\mu}) + \zeta_3 \log(\tfrac{2}{\epsilon}) \\
& \qquad \qquad + \zeta_3(1 + \zeta_2) \frac{NT}{\beta} \bigl[\mathcal{K}(\nu_3, \mu) - \mathcal{K}(\overline{\mu}, \mu)\bigr], \\
& \zeta_4 \frac{NT}{\beta} \nu_3 \bigl[\mathcal{K}(\rho_4, \overline{\pi})\bigr] \leq \zeta_4 \gamma \nu_3 (\rho_4 - \overline{\pi})(r) \\
& \quad + \zeta_4 C \nu_3 (\rho_4 \otimes \overline{\pi})(m') + \zeta_4 \mathcal{K}(\nu_3, \overline{\mu}) + \zeta_4 \nu_3 \bigl[\mathcal{K}(\rho_4, \overline{\pi})\bigr] + \zeta_4 \log(\tfrac{2}{\epsilon}) \\
& \qquad \qquad + \zeta_4 \frac{NT}{\beta} \nu_3 \bigl[\mathcal{K}(\rho_4, \pi) - \mathcal{K}(\overline{\pi}, \pi)\bigr], \\
& \zeta_5 \frac{NT}{\beta} \overline{\mu}\bigl[\mathcal{K}(\rho_5, \overline{\pi})\bigr] \leq \zeta_5 \gamma \overline{\mu}(\rho_5 - \overline{\pi})(r) + \zeta_5 C \overline{\mu}(\rho_5 \otimes \overline{\pi})(m') \\
& \qquad + \zeta_5 \overline{\mu}\bigl[\mathcal{K}(\rho_5, \overline{\pi})\bigr] + \zeta_5 \log(\tfrac{2}{\epsilon}) + \zeta_5 \frac{NT}{\beta} \overline{\mu}\bigl[\mathcal{K}(\rho_5, \pi) - \mathcal{K}(\overline{\pi}, \pi)\bigr].
\end{aligned}
$$

Adding these six inequalities and assuming that $\zeta_4 \leq \zeta_3 \bigl[(1 + \zeta_2) \frac{NT}{\beta} - 1\bigr]$, we find

$$
\begin{aligned}
& - N \log\bigl[1 - T(\nu\rho - \overline{\mu}\,\overline{\pi})(R)\bigr] - \beta(\nu\rho - \overline{\mu}\,\overline{\pi})(R) \\
& \quad \leq - N \log\bigl[1 - T(\nu\rho - \overline{\mu}\,\overline{\pi})(R)\bigr] - \beta(\nu\rho - \overline{\mu}\,\overline{\pi})(R) \\
& \qquad + \zeta_1 \bigl(\tfrac{NT}{\beta} - 1\bigr) \overline{\mu}\bigl[\mathcal{K}(\rho_1, \overline{\pi})\bigr] + \zeta_2 \bigl(\tfrac{NT}{\beta} - 1\bigr) \nu\bigl[\mathcal{K}(\rho_2, \overline{\pi})\bigr] \\
& \qquad \qquad + \bigl[\zeta_3(1 + \zeta_2) \tfrac{NT}{\beta} - \zeta_3 - \zeta_4\bigr] \mathcal{K}(\nu_3, \overline{\mu}) \\
& \qquad \qquad + \zeta_4 \bigl(\tfrac{NT}{\beta} - 1\bigr) \nu_3 \bigl[\mathcal{K}(\rho_4, \overline{\pi})\bigr] + \zeta_5 \bigl(\tfrac{NT}{\beta} - 1\bigr) \overline{\mu}\bigl[\mathcal{K}(\rho_5, \overline{\pi})\bigr] \\
& \quad \leq \gamma(\nu\rho - \overline{\mu}\,\overline{\pi})(r) + \zeta_1 \gamma \overline{\mu}(\rho_1 - \overline{\pi})(r) + \zeta_2 \gamma \nu(\rho_2 - \overline{\pi})(r)
\end{aligned}
$$

$$
\begin{aligned}
&+ \zeta_3 \gamma (\nu_3 - \overline{\mu}) \overline{\pi}(r) + \zeta_4 \gamma \nu_3 (\rho_4 - \overline{\pi})(r) + \zeta_5 \gamma \overline{\mu} (\rho_5 - \overline{\pi})(r) \\
&\qquad + C \bigl[(\nu\rho) \otimes (\overline{\mu}\,\overline{\pi}) + \zeta_1 \overline{\mu} (\rho_1 \otimes \overline{\pi}) + \zeta_2 \nu (\rho_2 \otimes \overline{\pi}) \\
&+ \zeta_3 (\nu_3 \overline{\pi}) \otimes (\nu_3 \rho_1) + \zeta_3 (\nu_3 \rho_1) \otimes (\overline{\mu}\,\overline{\pi}) \\
&\qquad + \zeta_4 \nu_3 (\rho_4 \otimes \overline{\pi}) + \zeta_5 \overline{\mu} (\rho_5 \otimes \overline{\pi}) \bigr](m') \\
&+ (1 + \zeta_2) \bigl[\mathcal{K}(\nu, \mu) - \mathcal{K}(\overline{\mu}, \mu) \bigr] + \nu \bigl[\mathcal{K}(\rho, \pi) - \mathcal{K}(\overline{\pi}, \pi) \bigr] \\
&\qquad + \zeta_1 \tfrac{NT}{\beta} \overline{\mu} \bigl[\mathcal{K}(\rho_1, \pi) - \mathcal{K}(\overline{\pi}, \pi) \bigr] + \zeta_2 \tfrac{NT}{\beta} \nu \bigl[\mathcal{K}(\rho_2, \pi) - \mathcal{K}(\overline{\pi}, \pi) \bigr] \\
&+ \zeta_3 (1 + \zeta_2) \tfrac{NT}{\beta} \bigl[\mathcal{K}(\nu_3, \mu) - \mathcal{K}(\overline{\mu}, \mu) \bigr] + \zeta_4 \tfrac{NT}{\beta} \nu_3 \bigl[\mathcal{K}(\rho_4, \pi) - \mathcal{K}(\overline{\pi}, \pi) \bigr] \\
&\qquad + \zeta_5 \tfrac{NT}{\beta} \overline{\mu} \bigl[\mathcal{K}(\rho_5, \pi) - \mathcal{K}(\overline{\pi}, \pi) \bigr] + (1 + \zeta_1 + \zeta_2 + \zeta_3 + \zeta_4 + \zeta_5) \log(\tfrac{2}{\epsilon}).
\end{aligned}
$$

Let us now apply to $\overline{\pi}$ (we shall later do the same with $\overline{\mu}$) the following inequalities, holding for any random functions of the sample and the parameters $h : \Omega \times \Theta \to \mathbb{R}$ and $g : \Omega \times \Theta \to \mathbb{R}$,

$$
\begin{aligned}
\overline{\pi}(g - h) - \mathcal{K}(\overline{\pi}, \pi) &\leq \sup_{\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)} \rho(g - h) - \mathcal{K}(\rho, \pi) \\
&= \log \bigl\{ \pi \bigl[\exp(g - h) \bigr] \bigr\} \\
&= \log \bigl\{ \pi \bigl[\exp(-h) \bigr] \bigr\} + \log \bigl\{ \pi_{\exp(-h)} \bigl[\exp(g) \bigr] \bigr\} \\
&= - \pi_{\exp(-h)}(h) - \mathcal{K}(\pi_{\exp(-h)}, \pi) + \log \bigl\{ \pi_{\exp(-h)} \bigl[\exp(g) \bigr] \bigr\}.
\end{aligned}
$$

When h and g are observable, and h is not too far from $\beta r \simeq \beta R$, this gives a way to replace $\overline{\pi}$ with some satisfactory empirical approximation. We will apply this method, choosing ρ_1 and ρ_5 such that $\overline{\mu}\,\overline{\pi}$ is replaced either with $\overline{\mu}\rho_1$, when it comes from the first two inequalities or with $\overline{\mu}\rho_5$ otherwise, choosing ρ_2 such that $\nu\overline{\pi}$ is replaced with $\nu\rho_2$ and ρ_4 such that $\nu_3\overline{\pi}$ is replaced with $\nu_3\rho_4$. We will do so because it leads to a lot of helpful cancellations. For those to happen, we need to choose $\rho_i = \pi_{\exp(-\lambda_i r)}$, $i = 1, 2, 4$, where λ_1, λ_2 and λ_4 are such that

$$
\begin{aligned}
(1 + \zeta_1)\gamma &= \zeta_1 \tfrac{NT}{\beta} \lambda_1, \\
\zeta_2 \gamma &= \bigl(1 + \zeta_2 \tfrac{NT}{\beta} \bigr) \lambda_2, \\
(\zeta_4 - \zeta_3)\gamma &= \zeta_4 \frac{NT}{\beta} \lambda_4, \\
\zeta_3 \gamma &= \zeta_5 \tfrac{NT}{\beta} \lambda_5,
\end{aligned}
$$

and to assume that $\zeta_4 > \zeta_3$. We obtain that with $\mathbb{P}$ probability at least $1 - \epsilon$,

$$
- N \log \bigl[1 - T(\mu\rho - \overline{\mu}\,\overline{\pi})(R) \bigr] - \beta(\nu\rho - \overline{\mu}\,\overline{\pi})(R)
$$

$$
\begin{aligned}
&\leq \gamma(\nu\rho - \overline{\mu}\,\rho_1)(r) + \zeta_3\gamma(\nu_3\rho_4 - \overline{\mu}\rho_5)(r) \\
&\quad + \zeta_1 \tfrac{NT}{\beta}\overline{\mu}\biggl\{\log\biggl[\rho_1\biggl\{\exp\biggl[C\tfrac{\beta}{NT\zeta_1}\bigl[\nu\rho + \zeta_1\rho_1\bigr](m')\biggr]\biggr\}\biggr]\biggr\} \\
&\quad + \bigl(1 + \zeta_2\tfrac{NT}{\beta}\bigr)\nu\biggl\{\log\biggl\{\rho_2\biggl\{\exp\biggl[\tfrac{C}{1+\zeta_2\frac{NT}{\beta}}\zeta_2\rho_2(m')\biggr]\biggr\}\biggr]\biggr\} \\
&\quad + \zeta_4\tfrac{NT}{\beta}\nu_3\biggl\{\log\biggl[\rho_4\biggl\{\exp\biggl[C\tfrac{\beta}{NT\zeta_4}\bigl[\zeta_3\nu_3\rho_1 + \zeta_4\rho_4\bigr](m')\biggr]\biggr\}\biggr]\biggr\} \\
&\quad + \zeta_5\tfrac{NT}{\beta}\overline{\mu}\biggl\{\log\biggl[\rho_5\biggl\{\exp\biggl[C\tfrac{\beta}{NT\zeta_5}\bigl[\zeta_3\nu_3\rho_1 + \zeta_5\rho_5\bigr](m')\biggr]\biggr\}\biggr]\biggr\} \\
&\quad + (1+\zeta_2)\bigl[\mathcal{K}(\nu,\mu) - \mathcal{K}(\overline{\mu},\mu)\bigr] + \nu\bigl[\mathcal{K}(\rho,\pi) - \mathcal{K}(\rho_2,\pi)\bigr] \\
&\quad + \zeta_3(1+\zeta_2)\tfrac{NT}{\beta}\bigl[\mathcal{K}(\nu_3,\mu) - \mathcal{K}(\overline{\mu},\mu)\bigr] \\
&\quad + \left(1 + \sum_{i=1}^{5}\zeta_i\right)\log\bigl(\tfrac{2}{\epsilon}\bigr).
\end{aligned}
$$

In order to obtain more cancellations while replacing $\overline{\mu}$ by some posterior distribution, we will choose the constants such that $\lambda_5 = \lambda_4$, which can be done by choosing

$$
\zeta_5 = \frac{\zeta_3\zeta_4}{\zeta_4 - \zeta_3}.
$$

We can now replace $\overline{\mu}$ with $\mu_{\exp -\xi_1\rho_1(r) - \xi_4\rho_4(r)}$, where

$$
\begin{aligned}
\xi_1 &= \frac{\gamma}{(1+\zeta_2)\bigl(1 + \frac{NT}{\beta}\zeta_3\bigr)}, \\
\xi_4 &= \frac{\gamma\zeta_3}{(1+\zeta_2)\bigl(1 + \frac{NT}{\beta}\zeta_3\bigr)}.
\end{aligned}
$$

Choosing moreover $\nu_3 = \mu_{\exp -\xi_1\rho_1(r) - \xi_4\rho_4(r)}$, to induce some more cancellations, we get

Theorem 1.59. *For any positive real constants satisfying the above mentioned constraints, with $\mathbb{P}$ probability at least $1-\epsilon$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$ and any conditional posterior distribution $\rho : \Omega \times M \to \mathcal{M}_+^1(\Theta)$,*

$$
- N\log\bigl[1 - T(\nu\rho - \overline{\mu}\,\overline{\pi})(R)\bigr] - \beta(\nu\rho - \overline{\mu}\,\overline{\pi})(R) \leq B(\nu,\rho,\beta),
$$

$$
\begin{aligned}
&\textit{where } B(\nu,\rho,\beta) \overset{\text{def}}{=} \gamma(\nu\rho - \nu_3\rho_1)(r) \\
&\quad + (1+\zeta_2)\bigl(1 + \tfrac{NT}{\beta}\zeta_3\bigr) \\
&\quad\times \log\Biggl\{ \nu_3 \Biggl[\rho_1 \biggl\{ \exp\Bigl[C \tfrac{\beta}{NT\zeta_1} \bigl[\nu\rho + \zeta_1\rho_1\bigr](m') \Bigr] \biggr\}^{\frac{\zeta_1 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}} \\
&\qquad\qquad \times \rho_4 \biggl\{ \exp\Bigl[C \tfrac{\beta}{NT\zeta_5} \bigl[\zeta_3\nu_3\rho_1 + \zeta_5\rho_4\bigr](m') \Bigr] \biggr\}^{\frac{\zeta_5 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}} \Biggr] \Biggr\} \\
&\qquad + \bigl(1 + \zeta_2 \tfrac{NT}{\beta}\bigr)\nu \Biggl\{ \log\biggl\{ \rho_2 \biggl\{ \exp\Bigl[\tfrac{C}{1+\zeta_2\frac{NT}{\beta}} \zeta_2\rho_2(m') \Bigr] \biggr\} \biggr] \Biggr\} \\
&\qquad + \zeta_4 \tfrac{NT}{\beta}\nu_3 \Biggl\{ \log\biggl[\rho_4 \biggl\{ \exp\Bigl[C \tfrac{\beta}{NT\zeta_4} \bigl[\zeta_3\nu_3\rho_1 + \zeta_4\rho_4\bigr](m') \Bigr] \biggr\} \biggr] \Biggr\} \\
&\quad + (1+\zeta_2)\bigl[\mathcal{K}(\nu,\mu) - \mathcal{K}(\nu_3,\mu)\bigr] \\
&\qquad\qquad\qquad + \nu\bigl[\mathcal{K}(\rho,\pi) - \mathcal{K}(\rho_2,\pi)\bigr] + \Biggl(1 + \sum_{i=1}^{5}\zeta_i\Biggr)\log\bigl(\tfrac{2}{\epsilon}\bigr).
\end{aligned}
$$

This theorem can be used to find the largest value $\widehat{\beta}(\nu\rho)$ of β such that $B(\nu,\rho,\beta) \leq 0$, thus providing an estimator for $\beta(\nu\rho)$ defined as $\nu\rho(R) = \overline{\mu}_{\beta(\nu\rho)}\overline{\pi}_{\beta(\nu\rho)}(R)$, where we have mentioned explicitely the dependence of $\overline{\mu}$ and $\overline{\pi}$ in β, the constant ζ_2 staying fixed. The posterior distribution $\nu\rho$ may then be chosen to maximize $\widehat{\beta}(\nu\rho)$ within some manageable subset of posterior distributions $\mathcal{P}$, thus gaining the assurance that $\nu\rho(R) \leq \overline{\mu}_{\widehat{\beta}(\nu\rho)}\overline{\pi}_{\widehat{\beta}(\nu\rho)}(R)$, with the largest parameter $\widehat{\beta}(\nu\rho)$ that this approach can provide. Maximizing $\widehat{\beta}(\nu\rho)$ is supported by the fact that $\lim_{\beta\to+\infty}\overline{\mu}_\beta\overline{\pi}_\beta(R) = \operatorname{ess\,inf}_{\mu\pi} R$. Anyhow, there is no assurance (to our knowledge) that $\beta \mapsto \overline{\mu}_\beta\overline{\pi}_\beta(R)$ will be a decreasing function of β all the way, although this may be expected to be the case in many practical situations.

We can make the bound more explicit in several ways. One point of view is to put forward the optimal values of ρ and ν. We can thus remark that

$$
\begin{aligned}
&\nu\bigl[\gamma\rho(r) + \mathcal{K}(\rho,\pi) - \mathcal{K}(\rho_2,\pi)\bigr] + (1+\zeta_2)\mathcal{K}(\nu,\mu) \\
&\quad = \nu\Biggl[\mathcal{K}\bigl[\rho,\pi_{\exp(-\gamma r)}\bigr] + \lambda_2\rho_2(r) + \int_{\lambda^2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\Biggr] + (1+\zeta_2)\mathcal{K}(\nu,\mu) \\
&\quad = \nu\bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\gamma r)}\bigr]\bigr\} + (1+\zeta_2)\mathcal{K}\bigl[\nu,\mu_{\exp\left(-\frac{\lambda_2\rho_2(r)}{1+\zeta_2} - \frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\right)}\bigr] \\
&\qquad - (1+\zeta_2)\log\Biggl\{\mu\Biggl[\exp\biggl\{-\frac{\lambda_2}{1+\zeta_2}\rho_2(r) - \frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\biggr\}\Biggr]\Biggr\}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
B(\nu,\rho,\beta) = (1+\zeta_2)\Bigl[&\xi_1\nu_3\rho_1(r) + \xi_4\nu_3\rho_4(r) \\
&+ \log\bigl\{\mu\bigl[\exp\bigl(-\xi_1\rho_1(r) - \xi_4\rho_4(r)\bigr)\bigr]\bigr\}\Bigr] \\
- (1+\zeta_2)\log&\left\{\mu\left[\exp\left\{-\frac{\lambda_2}{1+\zeta_2}\rho_2(r) - \frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\right\}\right]\right\} \\
- \gamma\nu_3\rho_1(r) &+ (1+\zeta_2)\bigl(1+\tfrac{NT}{\beta}\zeta_3\bigr) \\
&\times\log\left\{\nu_3\left[\rho_1\left\{\exp\left[C\tfrac{\beta}{NT\zeta_1}\bigl[\nu\rho+\zeta_1\rho_1\bigr](m')\right]\right\}^{\frac{\zeta_1 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}}\right.\right. \\
&\qquad\left.\left.\times\,\rho_4\left\{\exp\left[C\tfrac{\beta}{NT\zeta_5}\bigl[\zeta_3\nu_3\rho_1+\zeta_5\rho_4\bigr](m')\right]\right\}^{\frac{\zeta_5 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}}\right]\right\} \\
&+ \bigl(1+\zeta_2\tfrac{NT}{\beta}\bigr)\nu\left\{\log\left\{\rho_2\left\{\exp\left[\tfrac{C}{1+\zeta_2\frac{NT}{\beta}}\zeta_2\rho_2(m')\right]\right\}\right]\right\} \\
&+ \zeta_4\tfrac{NT}{\beta}\nu_3\left\{\log\left[\rho_4\left\{\exp\left[C\tfrac{\beta}{NT\zeta_4}\bigl[\zeta_3\nu_3\rho_1+\zeta_4\rho_4\bigr](m')\right]\right\}\right]\right\} \\
+ \nu\bigl\{\mathcal{K}\bigl[\rho,\pi_{\exp(-\gamma r)}\bigr]\bigr\}& \\
&+ (1+\zeta_2)\mathcal{K}\Bigl[\nu,\mu_{\exp\bigl(-\frac{\lambda_2\rho_2(r)}{1+\zeta_2}-\frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\bigr)}\Bigr] \\
&\qquad\qquad + \left(1+\sum_{i=1}^{5}\zeta_i\right)\log\bigl(\tfrac{2}{\epsilon}\bigr).
\end{aligned}
$$

This formula is better understood when thinking about the following upper bound for the two first lines in the expression of $B(\nu,\rho,\beta)$:

$$
\begin{aligned}
(1+\zeta_2)\Bigl[\xi_1\nu_3\rho_1(r) &+ \xi_4\nu_3\rho_4(r) + \log\bigl\{\mu\bigl[\exp\bigl(-\xi_1\rho_1(r) - \xi_4\rho_4(r)\bigr)\bigr]\bigr\}\Bigr] \\
- (1+\zeta_2)\log&\left\{\mu\left[\exp\left\{-\frac{\lambda_2}{1+\zeta_2}\rho_2(r)\right.\right.\right. \\
&\left.\left.\left. - \frac{1}{1+\zeta_2}\int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha\right\}\right]\right\} - \gamma\nu_3\rho_1(r) \\
&\leq \nu_3\left[\lambda_2\rho_2(r) + \int_{\lambda_2}^{\gamma}\pi_{\exp(-\alpha r)}(r)d\alpha - \gamma\rho_1(r)\right].
\end{aligned}
$$

Another approach to understanding Theorem 1.59 is to put forward $\rho_0 = \pi_{\exp(-\lambda_0 r)}$, for some positive real constant $\lambda_0 < \gamma$, noticing that

$$\nu\bigl[\mathcal{K}(\rho_0, \pi) - \mathcal{K}(\rho_2, \pi)\bigr] = \lambda_0 \nu(\rho_2 - \rho_0)(r) - \nu\bigl[\mathcal{K}(\rho_2, \rho_0)\bigr].$$

Thus

$$\begin{aligned} B(\nu, \rho_0, \beta) \leq \nu_3\bigl[(\gamma - \lambda_0)(\rho_0 - \rho_1)(r) + \lambda_0(\rho_2 - \rho_1)(r)\bigr] \\ + (1+\zeta_2)\bigl(1 + \tfrac{NT}{\beta}\zeta_3\bigr) \\ \times \log\Biggl\{\nu_3\Biggl[\rho_1\Biggl\{\exp\Biggl[C\tfrac{\beta}{NT\zeta_1}\bigl[\nu\rho_0 + \zeta_1\rho_1\bigr](m')\Biggr]\Biggr\}^{\frac{\zeta_1 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}} \\ \times \rho_4\Biggl\{\exp\Biggl[C\tfrac{\beta}{NT\zeta_5}\bigl[\zeta_3\nu_3\rho_1 + \zeta_5\rho_4\bigr](m')\Biggr]\Biggr\}^{\frac{\zeta_5 NT}{\beta(1+\zeta_2)(1+\frac{NT}{\beta}\zeta_3)}}\Biggr]\Biggr\} \\ + \bigl(1 + \zeta_2\tfrac{NT}{\beta}\bigr)\nu\Biggl\{\log\Biggl\{\rho_2\Biggl\{\exp\Biggl[\tfrac{C}{1+\zeta_2\frac{NT}{\beta}}\zeta_2\rho_2(m')\Biggr]\Biggr\}\Biggr]\Biggr\} \\ + \zeta_4\tfrac{NT}{\beta}\nu_3\Biggl\{\log\Biggl[\rho_4\Biggl\{\exp\Biggl[C\tfrac{\beta}{NT\zeta_4}\bigl[\zeta_3\nu_3\rho_1 + \zeta_4\rho_4\bigr](m')\Biggr]\Biggr\}\Biggr]\Biggr\} \\ + (1+\zeta_2)\mathcal{K}\Bigl[\nu, \mu_{\exp\left(-\frac{(\gamma-\lambda_0)\rho_0(r) + \lambda_0\rho_2(r)}{1+\zeta_2}\right)}\Bigr] \\ - \nu\bigl[\mathcal{K}(\rho_2, \rho_0)\bigr] + \Biggl(1 + \sum_{i=1}^{5}\zeta_i\Biggr)\log\bigl(\tfrac{2}{\epsilon}\bigr). \end{aligned}$$

In the case when we want to select a single model $\widehat{m}(\omega)$, and therefore to set $\nu = \delta_{\widehat{m}}$, the previous inequality engages us to take

$$\widehat{m} \in \arg\min_{m \in M} (\gamma - \lambda_0)\rho_0(m, r) + \lambda_0\rho_2(m, r).$$

In parametric situations where $\pi_{\exp(-\lambda r)}(r) \simeq r^\star(m) + \frac{d_e(m)}{\lambda}$, we get

$$(\gamma - \lambda_0)\rho_0(m, r) - \lambda_0\rho_2(m, r) \simeq \gamma\bigl[r^\star(m) + d_e(m)\bigl(\tfrac{1}{\lambda_0} + \tfrac{\lambda_0 - \lambda_2}{\gamma\lambda_2}\bigr)\bigr],$$

resulting in a linear penalization of the empirical dimension of the models.

1.5.8. Analysis of the two step relative bound. We will not state a formal result, but will neverless give some hints about how to establish one. We should start from Theorem 1.25, which gives a deterministic variance term. From Theorem 1.25, after a change of prior distribution, we obtain for any positive constants α_1 and α_2, any prior distributions $\widetilde{\mu}_1$ and $\widetilde{\mu}_2 \in \mathcal{M}_+^1(M)$, for any prior conditional distributions $\widetilde{\pi}_1$ and $\widetilde{\pi}_2 : M \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \eta$, for any posterior distributions $\nu_1\rho_1$ and $\nu_2\rho_2$,

$$\begin{aligned}\alpha_1(\nu_1\rho_1 - \nu_2\rho_2)(R) &\leq \alpha_2(\nu_1\rho_1 - \nu_2\rho_2)(r)\\ &\quad + \mathcal{K}\bigl[(\nu_1\rho_1)\otimes(\nu_2\rho_2), (\widetilde{\mu}_1\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_2\,\widetilde{\pi}_2)\bigr]\\ &\quad + \log\Bigl\{(\widetilde{\mu}_1\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_2\,\widetilde{\pi}_2)\Bigl[\exp\bigl\{-\alpha_2\Psi_{\frac{\alpha_2}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} - \log(\eta).\end{aligned}$$

Applying this to $\alpha_1 = 0$, we get that

$$\begin{aligned}(\nu\rho - \nu_3\rho_1)(r) \leq \frac{1}{\alpha_2}\Bigl[&\mathcal{K}\bigl[(\nu\rho)\otimes(\nu_3\rho_1), (\widetilde{\mu}\,\widetilde{\pi})\otimes(\widetilde{\mu}_3\,\widetilde{\pi}_1)\bigr]\\ &+ \log\Bigl\{(\widetilde{\mu}\,\widetilde{\nu})\otimes(\widetilde{\mu}_3\,\widetilde{\pi}_1)\Bigl[\exp\bigl\{\alpha_2\Psi_{-\frac{\alpha_2}{N}}(R', M')\bigr\}\Bigr]\Bigr\} - \log(\eta)\Bigr].\end{aligned}$$

In the same way, to bound quantities of the form

$$\begin{aligned}&\log\Biggl\{\nu_3\Biggl[\rho_1\Biggl\{\exp\Bigl[C_1(\nu\rho + \zeta_1\rho_1)(m')\Bigr]\Biggr\}^{p_1}\\ &\qquad\times\rho_4\Biggl\{\exp\Bigl[C_2\bigl[\zeta_3\nu_3\rho_1 + \zeta_5\rho_4\bigr](m')\Bigr]\Biggr\}^{p_2}\Biggr]\Biggr\}\\ &= \sup_{\nu_5}\Biggl\{p_1\sup_{\rho_5}\Bigl\{C_1\bigl[(\nu\rho)\otimes(\nu_5\rho_5) + \zeta_1\nu_5(\rho_1\otimes\rho_5)\bigr](m') - \mathcal{K}(\rho_5,\rho_1)\Bigr\}\\ &\quad + p_2\sup_{\rho_6}\Bigl\{C_2\bigl[\zeta_3(\nu_3\rho_1)\otimes(\nu_5\rho_6)\\ &\qquad\qquad + \zeta_5\nu_5(\rho_4\otimes\rho_6)\bigr](m') - \mathcal{K}(\rho_6,\rho_4)\Bigr\} - \mathcal{K}(\nu_5,\nu_3)\Biggr\},\end{aligned}$$

where C_1, C_2, p_1 and p_2 are positive constants, and similar terms, we need to use inequalities of the type: for any prior distributions $\widetilde{\mu}_i\,\widetilde{\pi}_i$, $i = 1, 2$, with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distributions $\nu_i\rho_i$, $i = 1, 2$,

$$\begin{aligned}\alpha_3(\nu_1\rho_1)\otimes(\nu_2\rho_2)(m') &\leq \log\Bigl\{(\widetilde{\mu}_1\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_2\,\widetilde{\pi}_2)\exp\Bigl[\alpha_3\Phi_{\frac{-\alpha_3}{N}}(M')\Bigr]\Bigr\}\\ &\quad + \mathcal{K}\bigl[(\nu_1\rho_1)\otimes(\nu_2\rho_2), (\widetilde{\mu}_1\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_2\,\widetilde{\pi}_2)\bigr] - \log(\eta).\end{aligned}$$

We need also the variant: with $\mathbb{P}$ probability at least $1-\eta$, for any posterior distribution $\nu_1 : \Omega \rightarrow \mathcal{M}_+^1(M)$ and any conditional posterior distributions $\rho_1, \rho_2 : \Omega\times M \rightarrow \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned}\alpha_3\nu_1(\rho_1\otimes\rho_2)(m') &\leq \log\Bigl\{\widetilde{\mu}_1\bigl(\widetilde{\pi}_1\otimes\widetilde{\pi}_2\bigr)\exp\Bigl[\alpha_3\Phi_{-\frac{\alpha_3}{N}}(M')\Bigr]\Bigr\}\\ &\quad + \mathcal{K}(\nu_1, \widetilde{\mu}_1) + \nu_1\bigl\{\mathcal{K}\bigl[\rho_1\otimes\rho_2, \widetilde{\pi}_1\otimes\widetilde{\pi}_2\bigr]\bigr\} - \log(\eta).\end{aligned}$$

We deduce that

$$
\begin{aligned}
&\log\biggl\{\nu_3\biggl[\rho_1\biggl\{\exp\Bigl[C_1(\nu\rho+\zeta_1\rho_1)(m')\Bigr]\biggr\}^{p_1} \\
&\qquad\qquad\qquad \times \rho_4\biggl\{\exp\Bigl[C_2\bigl[\zeta_3\nu_3\rho_1+\zeta_5\rho_4\bigr](m')\Bigr]\biggr\}^{p_2}\biggr]\biggr\} \\
&\leq \sup_{\nu_5}\biggl\{p_1\sup_{\rho_5}\biggl[\frac{C_1}{\alpha_3}\biggl\{\log\Bigl\{(\widetilde{\mu}\,\widetilde{\pi})\otimes(\widetilde{\mu}_5\,\widetilde{\pi}_5)\exp\Bigl[\alpha_3\Phi_{-\frac{\alpha_3}{N}}(M')\Bigr]\Bigr\} \\
&\qquad + \mathcal{K}\bigl[(\nu\rho)\otimes(\nu_5\rho_5),(\widetilde{\mu}\,\widetilde{\pi}\otimes(\widetilde{\mu}_5\,\widetilde{\pi}_5)\bigr]+\log(\tfrac{2}{\eta}) \\
&\qquad + \zeta_1\biggl[\log\Bigl\{\widetilde{\mu}_5\bigl(\widetilde{\pi}_1\otimes\widetilde{\pi}_5\bigr)\exp\Bigl[\alpha_3\Phi_{-\frac{\alpha_3}{N}}(M')\Bigr]\Bigr\} \\
&+\mathcal{K}(\nu_5,\widetilde{\mu}_5)+\nu_5\bigl\{\mathcal{K}\bigl[\rho_1\otimes\rho_5,\widetilde{\pi}_1\otimes\widetilde{\pi}_5\bigr]\bigr\}+\log\bigl(\tfrac{2}{\eta}\bigr)\biggr]\biggr\}-\mathcal{K}(\rho_5,\rho_1)\biggr] \\
&\quad + p_2\sup_{\rho_6}\biggl[\frac{C_1}{\alpha_3}\biggl\{\log\Bigl\{(\widetilde{\mu}_3\,\widetilde{\pi}_1)\otimes(\widetilde{\mu}_5\,\widetilde{\pi}_6)\exp\Bigl[\alpha_3\Phi_{-\frac{\alpha_3}{N}}(M')\Bigr]\Bigr\} \\
&\qquad + \mathcal{K}\bigl[(\nu_3\rho_1)\otimes(\nu_5\rho_6),(\widetilde{\mu}_3\,\widetilde{\pi}_1\otimes(\widetilde{\mu}_5\,\widetilde{\pi}_6)\bigr]+\log(\tfrac{2}{\eta}) \\
&\qquad + \zeta_1\biggl[\log\Bigl\{\widetilde{\mu}_5\bigl(\widetilde{\pi}_4\otimes\widetilde{\pi}_6\bigr)\exp\Bigl[\alpha_3\Phi_{-\frac{\alpha_3}{N}}(M')\Bigr]\Bigr\} \\
&\qquad\quad +\mathcal{K}(\nu_5,\widetilde{\mu}_5)+\nu_5\bigl\{\mathcal{K}\bigl[\rho_4\otimes\rho_6,\widetilde{\pi}_4\otimes\widetilde{\pi}_6\bigr]\bigr\}+\log\bigl(\tfrac{2}{\eta}\bigr)\biggr]\biggr\} \\
&\qquad\qquad\qquad\qquad - \mathcal{K}(\rho_6,\rho_4)\biggr]-\mathcal{K}(\nu_5,\nu_3)\biggr\}.
\end{aligned}
$$

We are then left with the need to bound entropy terms like $\mathcal{K}(\nu_3\rho_1,\widetilde{\mu}_3\widetilde{\pi}_1)$, where we have the choice of $\widetilde{\mu}_3$ and $\widetilde{\pi}_1$, to obtain a useful bound. As could be expected, we decompose it into

$$
\mathcal{K}(\nu_3\rho_1,\widetilde{\mu}_3\widetilde{\pi}_1)=\mathcal{K}(\nu_3,\widetilde{\mu}_3)+\nu_3\bigl[\mathcal{K}(\rho_1,\widetilde{\pi}_1)\bigr].
$$

Let us look after the second term first, choosing $\widetilde{\pi}_1=\pi_{\exp(-\beta_1 R)}$:

$$
\begin{aligned}
\nu_3\bigl[\mathcal{K}(\rho_1,\widetilde{\pi}_1)\bigr] &= \nu_3\bigl[\beta_1(\rho_1-\widetilde{\pi}_1)(R)+\mathcal{K}(\rho_1,\pi)-\mathcal{K}(\widetilde{\pi}_1,\pi)\bigr] \\
&\leq \frac{\beta_1}{\alpha_1}\Bigl[\alpha_2\nu_3(\rho_1-\widetilde{\pi}_1)(r)+\mathcal{K}(\nu_3,\widetilde{\mu}_3)+\nu_3\bigl[\mathcal{K}(\rho_1,\widetilde{\pi}_1)\bigr] \\
&\quad + \log\Bigl\{\widetilde{\mu}_3\bigl(\widetilde{\pi}_1^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\alpha_2\Psi_{\frac{\alpha_2}{N}}(R',M')+\alpha_1 R'\bigr\}\Bigr]\Bigr\}-\log(\eta)\Bigr]
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad + \nu_3\bigl[\mathcal{K}(\rho_1,\pi) - \mathcal{K}(\widetilde{\pi}_1,\pi)\bigr] \\
&\leq \frac{\beta_1}{\alpha_1}\biggl[\mathcal{K}(\nu_3,\widetilde{\mu}_3) + \nu_3\bigl[\mathcal{K}(\rho_1,\widetilde{\pi}_1)\bigr] \\
&\quad + \log\Bigl\{\widetilde{\mu}_3\bigl(\widetilde{\pi}_1^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\alpha_2\Psi_{\frac{\alpha_2}{N}}(R',M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} - \log(\eta)\biggr] \\
&\qquad\qquad\qquad\qquad + \nu_3\bigl\{\mathcal{K}\bigl[\rho_1, \pi_{\exp(-\frac{\beta_1\alpha_2}{\alpha_1}r)}\bigr]\bigr\}.
\end{aligned}
$$

Thus, when the constraint $\lambda_1 = \frac{\beta_1\alpha_2}{\alpha_1}$ is satisfied,

$$
\begin{aligned}
\nu_3\bigl[\mathcal{K}(\rho_1,\widetilde{\pi}_1)\bigr] \leq \Bigl(1 - \frac{\beta_1}{\alpha_1}\Bigr)^{-1}\frac{\beta_1}{\alpha_1}\biggl[\mathcal{K}(\nu_3,\widetilde{\mu}_3) \\
+ \log\Bigl\{\widetilde{\mu}_3\bigl(\widetilde{\pi}_1^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\alpha_2\Psi_{\frac{\alpha_2}{N}}(R',M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} - \log(\eta)\biggr].
\end{aligned}
$$

We can further specialize the constants, choosing $\alpha_1 = N\sinh(\frac{\alpha_2}{N})$, so that

$$
-\alpha_2\Psi_{\frac{\alpha_2}{N}}(R',M') + \alpha_1 R' \leq 2N\sinh\Bigl(\frac{\alpha_2}{2N}\Bigr)^2 M'.
$$

We can for instance choose $\alpha_2 = \gamma$, $\alpha_1 = N\sinh(\frac{\gamma}{N})$, and $\beta_1 = \lambda_1\frac{N}{\gamma}\sinh(\frac{\gamma}{N})$, leading to

Proposition 1.60. *With the notations of Theorem 1.59, the constants being set as explained above, putting $\widetilde{\pi}_1 = \pi_{\exp(-\lambda_1\frac{N}{\gamma}\sinh(\frac{\gamma}{N})R)}$, with $\mathbb{P}$ probability at least $1-\eta$,*

$$
\begin{aligned}
\nu_3\bigl[\mathcal{K}(\rho_1,\widetilde{\pi}_1)\bigr] \leq \Bigl(1 - \frac{\lambda_1}{\gamma}\Bigr)^{-1}\frac{\lambda_1}{\gamma}\biggl[\mathcal{K}(\nu_3,\widetilde{\mu}_3) \\
+ \log\Bigl\{\widetilde{\mu}_3\bigl(\widetilde{\pi}_1^{\otimes 2}\bigr)\Bigl[\exp\bigl\{2N\sinh(\tfrac{\gamma}{2N})^2 M'\bigr\}\Bigr]\Bigr\} - \log(\eta)\biggr].
\end{aligned}
$$

More generally

$$
\begin{aligned}
\nu_3\bigl[\mathcal{K}(\rho,\widetilde{\pi}_1)\bigr] \leq \Bigl(1 - \frac{\lambda_1}{\gamma}\Bigr)^{-1}\frac{\lambda_1}{\gamma}\biggl[\mathcal{K}(\nu_3,\widetilde{\mu}_3) \\
+ \log\Bigl\{\widetilde{\mu}_3\bigl(\widetilde{\pi}_1^{\otimes 2}\bigr)\Bigl[\exp\bigl\{2N\sinh(\tfrac{\gamma}{2N})^2 M'\bigr\}\Bigr]\Bigr\} - \log(\eta)\biggr] \\
+ \Bigl(1 - \frac{\lambda_1}{\gamma}\Bigr)^{-1}\nu_3\bigl[\mathcal{K}(\rho,\rho_1)\bigr].
\end{aligned}
$$

In a similar way, let us choose now $\widetilde{\mu}_3 = \mu_{\exp[-\alpha_3 \overline{\pi}(R)]}$. We can write

$$\begin{aligned}\mathcal{K}(\nu, \widetilde{\mu}_3) = \alpha_3(\nu - \widetilde{\mu}_3)\overline{\pi}(R) + \mathcal{K}(\nu, \mu) - \mathcal{K}(\widetilde{\mu}_3, \mu) \\ \leq \frac{\alpha_3}{\alpha_1}\Bigl[\alpha_2(\nu - \widetilde{\mu}_3)\overline{\pi}(r) + \mathcal{K}(\nu, \widetilde{\mu}_3) \\ + \log\Bigl\{(\widetilde{\mu}_3\overline{\pi}) \otimes (\widetilde{\mu}_3\overline{\pi})\Bigl[\exp\bigl\{-\alpha_2\Psi_{\frac{\alpha_2}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} - \log(\eta)\Bigr] \\ + \mathcal{K}(\nu, \mu) - \mathcal{K}(\widetilde{\mu}_3, \mu).\end{aligned}$$

Let us choose $\alpha_2 = \gamma$, $\alpha_1 = N\sinh(\frac{\gamma}{N})$, and let us add some other entropy inequalities to get rid of $\overline{\pi}$ in a suitable way, the approach of entropy compensation being quite the same as the one used to obtain the empirical bound of Theorem 1.59. This results with $\mathbb{P}$ probability at least $1 - \eta$ in

$$\begin{aligned}\Bigl(1 - \frac{\alpha_3}{\alpha_1}\Bigr)\mathcal{K}(\nu, \widetilde{\mu}_3) \leq \frac{\alpha_3}{\alpha_1}\Bigl[\gamma(\nu - \widetilde{\mu}_3)\overline{\pi}(r) \\ + \log\Bigl\{(\widetilde{\mu}_3\overline{\pi}) \otimes (\widetilde{\mu}_3\overline{\pi})\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} + \log(\tfrac{2}{\eta})\Bigr] \\ + \mathcal{K}(\nu, \mu) - \mathcal{K}(\widetilde{\mu}_3, \mu), \\ \zeta_6\Bigl(1 - \frac{\beta}{\alpha_1}\Bigr)\widetilde{\mu}_3\bigl[\mathcal{K}(\rho_6, \overline{\pi})\bigr] \leq \zeta_6\frac{\beta}{\alpha_1}\Bigl[\gamma\widetilde{\mu}_3(\rho_6 - \overline{\pi})(r) \\ + \log\Bigl\{\widetilde{\mu}_3\bigl(\overline{\pi}^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} + \log(\tfrac{2}{\eta})\Bigr] \\ + \zeta_6\widetilde{\mu}_3\bigl[\mathcal{K}(\rho_6, \pi) - \mathcal{K}(\overline{\pi}, \pi)\bigr], \\ \zeta_7\Bigl(1 - \frac{\beta}{\alpha_1}\Bigr)\widetilde{\mu}_3\bigl[\mathcal{K}(\rho_7, \overline{\pi})\bigr] \leq \zeta_7\frac{\beta}{\alpha_1}\Bigl[\gamma\widetilde{\mu}_3(\rho_7 - \overline{\pi})(r) \\ + \log\Bigl\{\widetilde{\mu}_3\bigl(\overline{\pi}^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} + \log(\tfrac{2}{\eta})\Bigr] \\ + \zeta_7\widetilde{\mu}_3\bigl[\mathcal{K}(\rho_7, \pi) - \mathcal{K}(\overline{\pi}, \pi)\bigr], \\ \zeta_8\Bigl(1 - \frac{\beta}{\alpha_1}\Bigr)\nu\bigl[\mathcal{K}(\rho_8, \overline{\pi})\bigr] \leq \zeta_8\frac{\beta}{\alpha_1}\Bigl[\gamma\nu(\rho_8 - \overline{\pi})(r) + \mathcal{K}(\nu, \widetilde{\mu}_3) \\ + \log\Bigl\{\widetilde{\mu}_3\bigl(\overline{\pi}^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} + \log(\tfrac{2}{\eta})\Bigr] \\ + \zeta_8\nu\bigl[\mathcal{K}(\rho_8, \pi) - \mathcal{K}(\overline{\pi}, \pi)\bigr], \\ \zeta_9\Bigl(1 - \frac{\beta}{\alpha_1}\Bigr)\nu\bigl[\mathcal{K}(\rho_9, \overline{\pi})\bigr] \leq \zeta_9\frac{\beta}{\alpha_1}\Bigl[\gamma\nu(\rho_9 - \overline{\pi})(r) + \mathcal{K}(\nu, \widetilde{\mu}_3)\end{aligned}$$

$$+ \log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\big\}\Big]\Big\} + \log(\tfrac{2}{\eta})\Big]$$
$$+ \zeta_9 \nu\big[\mathcal{K}(\rho_9, \pi) - \mathcal{K}(\overline{\pi}, \pi)\big],$$

where we have introduced a bunch of constants, assumed to be positive, that we will more precisely set to

$$\begin{aligned}
x_8 + x_9 &= 1,\\
(\zeta_6\beta + x_8\alpha_3)\frac{\gamma}{\alpha_1} &= \lambda_6,\\
(\zeta_7\beta + x_9\alpha_3)\frac{\gamma}{\alpha_1} &= \lambda_7,\\
(\zeta_8\beta - x_8\alpha_3)\frac{\gamma}{\alpha_1} &= \lambda_8,\\
(\zeta_9\beta - x_9\alpha_3)\frac{\gamma}{\alpha_1} &= \lambda_9.
\end{aligned}$$

We get with $\mathbb{P}$ probability at least $1-\eta$,

$$\begin{aligned}
&\Big(1 - \frac{\alpha_3}{\alpha_1} - (\zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\Big)\mathcal{K}(\nu, \widetilde{\mu}_3) \leq \\
&\qquad \frac{\alpha_3}{\alpha_1}\Big[\gamma\big[\nu(x_8\rho_8 + x_9\rho_9)(r) - \widetilde{\mu}_3(x_8\rho_6 + x_9\rho_7)(r)\big]\\
&\qquad + \frac{\alpha_3}{\alpha_1}\log\Big\{(\widetilde{\mu}_3\overline{\pi}) \otimes (\widetilde{\mu}_3\overline{\pi})\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\big\}\Big]\Big\}\\
&\quad + (\zeta_6 + \zeta_7 + \zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\log\Big\{\widetilde{\mu}_3\big(\overline{\pi}^{\otimes 2}\big)\Big[\exp\big\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\big\}\Big]\Big\}\\
&\qquad + \mathcal{K}(\nu, \mu) - \mathcal{K}(\widetilde{\mu}_3, \mu) + \Big(\frac{\alpha_3}{\alpha_1} + (\zeta_6 + \zeta_7 + \zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\Big)\log\big(\tfrac{2}{\eta}\big).
\end{aligned}$$

Let us choose the constants so that $\lambda_1 = \lambda_7 = \lambda_9$, $\lambda_4 = \lambda_6 = \lambda_8$, $\alpha_3 x_9 \frac{\gamma}{\alpha_1} = \xi_1$ and $\alpha_3 x_8 \frac{\gamma}{\alpha_1} = \xi_4$. This is done by setting

$$\begin{aligned}
x_8 &= \frac{\xi_4}{\xi_1 + \xi_4},\\
x_9 &= \frac{\xi_1}{\xi_1 + \xi_4},\\
\alpha_3 &= \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})(\xi_1 + \xi_4),\\
\zeta_6 &= \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\frac{(\lambda_4 - \xi_4)}{\beta},\\
\zeta_7 &= \tfrac{N}{\gamma}\sinh(\tfrac{\gamma}{N})\frac{(\lambda_1 - \xi_1)}{\beta},
\end{aligned}$$

$$\zeta_8 = \tfrac{N}{\gamma} \sinh(\tfrac{\gamma}{N}) \frac{(\lambda_4 + \xi_4)}{\beta},$$
$$\zeta_9 = \tfrac{N}{\gamma} \sinh(\tfrac{\gamma}{N}) \frac{(\lambda_1 + \xi_1)}{\beta}.$$

The inequality $\lambda_1 > \xi_1$ is always satisfied. The inequality $\lambda_4 > \xi_4$ is required for the above choice of constants, and will be satisfied for a suitable choice of ζ_3 and ζ_4.

Under these asumptions, we obtain with $\mathbb{P}$ probability at least $1 - \eta$

$$\begin{aligned}\Bigl(1 - \frac{\alpha_3}{\alpha_1} - (\zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\Bigr)\mathcal{K}(\nu, \widetilde{\mu}_3) \leq (\nu - \widetilde{\mu}_3)(\xi_1\rho_1 + \xi_4\rho_4)(r) \\ + \frac{\alpha_3}{\alpha_1}\log\Bigl\{(\widetilde{\mu}_3\overline{\pi}) \otimes (\widetilde{\mu}_3\overline{\pi})\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} \\ + (\zeta_6 + \zeta_7 + \zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\log\Bigl\{\widetilde{\mu}_3\bigl(\overline{\pi}^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} \\ + \mathcal{K}(\nu, \mu) - \mathcal{K}(\widetilde{\mu}_3, \mu) + \Bigl(\frac{\alpha_3}{\alpha_1} + (\zeta_6 + \zeta_7 + \zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\Bigr)\log\bigl(\tfrac{2}{\eta}\bigr).\end{aligned}$$

This proves

Proposition 1.61. *The constants being set as explained above, with $\mathbb{P}$ probability at least $1 - \eta$, for any posterior distribution $\nu : \Omega \to \mathcal{M}_+^1(M)$,*

$$\begin{aligned}\mathcal{K}(\nu, \widetilde{\mu}_3) \leq \Bigl(1 - \frac{\alpha_3}{\alpha_1} - (\zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\Bigr)^{-1}\Bigl[\mathcal{K}(\nu, \nu_3) \\ + \frac{\alpha_3}{\alpha_1}\log\Bigl\{(\widetilde{\mu}_3\overline{\pi}) \otimes (\widetilde{\mu}_3\overline{\pi})\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} \\ + (\zeta_6 + \zeta_7 + \zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\log\Bigl\{\widetilde{\mu}_3\bigl(\overline{\pi}^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} \\ + \Bigl(\frac{\alpha_3}{\alpha_1} + (\zeta_6 + \zeta_7 + \zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\Bigr)\log\bigl(\tfrac{2}{\eta}\bigr)\Bigr].\end{aligned}$$

Thus

$$\begin{aligned}\mathcal{K}(\nu_3\rho_1, \widetilde{\mu}_3\,\widetilde{\pi}_1) \leq \frac{1 + \bigl(1 - \frac{\lambda_1}{\gamma}\bigr)^{-1}\frac{\lambda_1}{\gamma}}{1 - \frac{\alpha_3}{\alpha_1} - (\zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}} \\ \times \Bigl[\frac{\alpha_3}{\alpha_1}\log\Bigl\{(\widetilde{\mu}_3\overline{\pi} \otimes (\widetilde{\mu}_3\overline{\pi})\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\} \\ + (\zeta_6 + \zeta_7 + \zeta_8 + \zeta_9)\frac{\beta}{\alpha_1}\log\Bigl\{\widetilde{\mu}_3\bigl(\overline{\pi}^{\otimes 2}\bigr)\Bigl[\exp\bigl\{-\gamma\Psi_{\frac{\gamma}{N}}(R', M') + \alpha_1 R'\bigr\}\Bigr]\Bigr\}\end{aligned}$$

$$+ \Big(\frac{\alpha_3}{\alpha_1} + (\zeta_6 + \zeta_7 + \zeta_8 + \zeta_9) \frac{\beta}{\alpha_1} \Big) \log\bigl(\tfrac{2}{\eta}\bigr) \Big]$$
$$+ \Big(1 - \frac{\lambda_1}{\gamma} \Big)^{-1} \frac{\lambda_1}{\gamma} \Big[\log \Big\{ \widetilde{\mu}_3 \bigl(\widetilde{\pi}_1^{\otimes 2} \bigr) \Big[\exp\bigl\{ 2N \sinh\bigl(\tfrac{\gamma}{2N}\bigr)^2 M' \bigr\} \Big] \Big\} - \log\bigl(\tfrac{2}{\eta}\bigr) \Big].$$

We will not go further, lest it may become tedious, but we hope we have given sufficient hints to state informally that the bound $B(\nu, \rho, \beta)$ of Theorem 1.59 is upper bounded with $\mathbb{P}$ probability close to one by a bound of the same flavour where the empirical quantities r and m' have been replaced with their expectations R and M'.

2. Transductive PAC-Bayesian learning

2.1. Basic inequalities. In this section the observed sample $(X_i, Y_i)_{i=1}^N$ will be supplemented with a *shadow sample* $(X_i, Y_i)_{i=N+1}^{(k+1)N}$. This point of view, called *transductive classification*, has been introduced by V. Vapnik. It may be justified in different ways.

On the practical side, one interest of the transductive setting is that it is often a lot easier to collect examples than it is to label them, so that it is not unreallistic to assume that we indeed have two training samples, one labelled and one unlabelled. It also covers the case when a batch of patterns is to be classified and we are allowed to observe the whole batch before issuing the classification.

On the mathematical side, considering a shadow sample proves technically fruitfull. Indeed, when introducing the VC entropy and VC dimension concepts, as well as when dealing with compression schemes, albeit the *inductive* setting is our final concern, the transductive setting is a useful detour. In this second scenario, intermediate technical results involving the shadow sample are integrated with respect to unobserved random variables in a second stage of the proofs.

Let us describe now the changes to be made to previous notations to adapt them to the transductive setting. The distribution $\mathbb{P}$ will be a probability measure on the canonical space $\Omega = (\mathcal{X} \times \mathcal{Y})^{(k+1)N}$, and $(X_i, Y_i)_{i=1}^{(k+1)N}$ will be the canonical process on this space (that is the coordinate process). Unless explicitely mentioned, the parameter k indicating the size of the shadow sample will remain fixed. Assuming the shadow sample size is a multiple of the training sample size is convenient without significantly restricting the generality. For a while, we will use a weaker assumption than independence,

assuming that $\mathbb{P}$ is *partially exchangeable*, since this is all what we need in the proofs.

Definition 2.1. For $i = 1, \dots, N$, let $\tau_i : \Omega \to \Omega$ be defined for any $\omega = (\omega_j)_{j=1}^{(k+1)N} \in \Omega$ by

$$\begin{cases} \tau_i(\omega)_{i+jN} = \omega_{i+(j-1)N}, & j = 1, \dots, k, \\ \tau_i(\omega)_i = \omega_{i+kN}, \\ \text{and } \tau_i(\omega)_{m+jN} = \omega_{m+jN}, & m \neq i, m = 1, \dots, N, j = 0, \dots k. \end{cases}$$

Clearly, if we arrange the $(k+1)N$ samples in a $N \times (k+1)$ array, τ_i performs a circular permutation of $k+1$ entries on the ith row, letting the other rows unchanged. Moreover, all the circular permutations of the ith row have the form τ_i^j, j ranging from 0 to k.

The probability distribution $\mathbb{P}$ is said to be partially exchangeable if for any $i = 1, \dots, N$, $\mathbb{P} \circ \tau_i^{-1} = \mathbb{P}$.

This means equivalently that for any bounded measurable function $h : \Omega \to \mathbb{R}$, $\mathbb{P}(h \circ \tau_i) = \mathbb{P}(h)$.

In the same way a function h defined on Ω will be said to be partially exchangeable if $h \circ \tau_i = h$ for any $i = 1, \dots, N$. Accordingly a posterior distribution $\rho : \Omega \to \mathcal{M}_+^1(\Theta, \mathcal{T})$ will be said to be partially exchangeable when $\rho(\omega, A) = \rho\bigl[\tau_i(\omega), A\bigr]$, for any $\omega \in \Omega$, any $i = 1, \dots, N$ and any $A \in \mathcal{T}$.

For any bounded measurable function h, let us define $T_i(h) = \frac{1}{k+1} \sum_{j=0}^{k} h \circ \tau_i^j$. Let $T(h) = T_N \circ \dots \circ T_1(h)$. For any partially exchangeable probability distribution $\mathbb{P}$, and for any bounded measurable function h, $\mathbb{P}\bigl[T(h)\bigr] = \mathbb{P}(h)$. Let us put

$$\sigma_i(\theta) = \mathbb{1}\bigl[f_\theta(X_i) \neq Y_i\bigr], \quad \text{indicating the success or failure of } f_\theta \text{ to predict } Y_i \text{ from } X_i,$$

$$r_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i(\theta), \quad \text{the empirical error rate of } f_\theta \text{ on the observed sample},$$

$$r_2(\theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \sigma_i(\theta), \quad \text{the error rate of } f_\theta \text{ on the shadow sample},$$

$$\overline{r}(\theta) = \frac{r_1(\theta) + k r_2(\theta)}{k+1} = \frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} \sigma_i(\theta), \quad \text{the global error rate of } f_\theta,$$

$$R_i(\theta) = \mathbb{P}\bigl[f_\theta(X_i) \neq Y_i\bigr], \qquad \text{the expected error rate of } f_\theta \text{ on the } i\text{th input,}$$

$$R(\theta) = \frac{1}{N}\sum_{i=1}^N R_i(\theta) = \mathbb{P}\bigl[r_1(\theta)\bigr] = \mathbb{P}\bigl[r_2(\theta)\bigr], \quad \text{the average expected}$$

error rate of f_θ on all inputs.

We will allow for posterior distributions $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ depending on the shadow sample. The most interesting ones will anyhow be independent of the shadow labels $Y_{N+1}, \dots, Y_{(k+1)N}$. We will be interested in the conditional expected error rate of the randomized classification rule described by ρ on the shadow sample, given the observed sample, which reads as $\mathbb{P}\bigl[\rho(r_2) | (X_i, Y_i)_{i=1}^N\bigr]$.

Let us comment on the case when $\mathbb{P}$ is invariant by any permutations of the rows, meaning that

$$\mathbb{P}\bigl[h(\omega \circ s)\bigr] = \mathbb{P}\bigl[h(\omega)\bigr] \text{ for all } s \in \mathfrak{S}(\{i + jN; j = 0, \dots, k\})$$

and all $i = 1, \dots, N$ (where $\mathfrak{S}(A)$ is the set of permutations of A, extended to $\{1, \dots, (k+1)N\}$ so as to be the identity outside of A). In this case, if ρ is invariant by permutations of the rows of the shadow sample, meaning that $\rho(\omega \circ s) = \rho(\omega) \in \mathcal{M}_+^1(\Theta)$, $s \in \mathfrak{S}(\{i + jN; j = 1, \dots, k\})$, $i = 1, \dots, N$, then $\mathbb{P}\bigl[\rho(r_2) | (X_i, Y_i)_{i=1}^N\bigr] = \frac{1}{N}\sum_{i=1}^N \mathbb{P}\bigl[\rho(\sigma_{i+N}) | (X_i, Y_i)_{i=1}^N\bigr]$, meaning that the expectation can be taken on a restricted shadow sample of the same size as the observed sample. If moreover the rows are equidistributed (meaning that their marginal distributions are equal), then

$$\mathbb{P}\bigl[\rho(r_2) | (X_i, Y_i)_{i=1}^N\bigr] = \mathbb{P}\bigl[\rho(\sigma_{N+1}) | (X_i, Y_i)_{i=1}^N\bigr].$$

This means that under these quite commonly fullfilled assumptions, the expectation can be taken on a single new object to be classified, our study thus covers the case when only one of the patterns from the shadow sample is to be labelled and one is interested in the expected error rate of this single labelling. Of course, in the case when $\mathbb{P}$ is i.i.d. and ρ depends only on the training sample $(X_i, Y_i)_{i=1}^N$, we fall back on the usual criterion of performance $\mathbb{P}\bigl[\rho(r_2) | (Z_i)_{i=1}^N\bigr] = \rho(R) = \rho(R_1)$.

Let us recall the notation $\Phi_a(p) = -a^{-1}\log\bigl\{1 - p\bigl[1 - \exp(-a)\bigr]\bigr\}$.

Using an obvious factorization, and considering for the moment a fixed value of θ and any partially exchangeable positive real measurable function $\lambda : \Omega \rightarrow \mathbb{R}_+$, we can compute the log Laplace transform of r_1 under T, which acts like a conditional probability distribution:

$$\log\Bigl\{T\bigl[\exp(-\lambda r_1)\bigr]\Bigr\} = \sum_{i=1}^N \log\Bigl\{T_i\bigl[\exp(-\tfrac{\lambda}{N}\sigma_i)\bigr]\Bigr\}$$

$$\leq N \log \biggl\{ \frac{1}{N} \sum_{i=1}^{N} T_i \Bigl[\exp\bigl(- \tfrac{\lambda}{N} \sigma_i \bigr) \Bigr] \biggr\} = - \lambda \Phi_{\frac{\lambda}{N}}(\overline{r}).$$

Remarking that $T \Bigl\{ \exp \Bigl[\lambda \bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1 \bigr] \Bigr] \Bigr\} = \exp\bigl[\lambda \Phi_{\frac{\lambda}{N}}(\overline{r}) \bigr] T \bigl[\exp(-\lambda r_1) \bigr]$ we obtain

Lemma 2.1. *For any $\theta \in \Theta$ and any partially exchangeable positive real measurable function $\lambda : \Omega \rightarrow \mathbb{R}_+$,*

$$T \Bigl\{ \exp \Bigl[\lambda \bigl\{ \Phi_{\frac{\lambda}{N}} \bigl[\overline{r}(\theta) \bigr] - r_1(\theta) \bigr\} \Bigr] \Bigr\} \leq 1.$$

We deduce from this lemma a result analogous to the inductive case:

Theorem 2.2. *For any partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \rightarrow \mathbb{R}_+$, for any partially exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P} \biggl\{ \exp \biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho \Bigl[\lambda \bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1 \bigr] \Bigr] - \mathcal{K}(\rho, \pi) \biggr] \biggr\} \leq 1.$$

The proof is deduced from the previous lemma, using the fact that π is partially exchangeable :

$$\begin{aligned}
&\mathbb{P} \biggl\{ \exp \biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho \Bigl[\lambda \bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1 \bigr] \Bigr] - \mathcal{K}(\rho, \pi) \biggr] \biggr\} \\
&\quad = \mathbb{P} \biggl\{ \pi \Bigl\{ \exp \Bigl[\lambda \bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1 \bigr] \Bigr] \Bigr\} \biggr\} = \mathbb{P} \biggl\{ T \pi \Bigl\{ \exp \Bigl[\lambda \bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1 \bigr] \Bigr] \Bigr\} \biggr\} \\
&\qquad = \mathbb{P} \biggl\{ \pi \Bigl\{ T \exp \Bigl[\lambda \bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}) - r_1 \bigr] \Bigr] \Bigr\} \biggr\} \leq 1.
\end{aligned}$$

Introducing in the same way

$$m'(\theta, \theta') = \frac{1}{N} \sum_{i=1}^{N} \Bigl| \mathbb{1} \bigl[f_\theta(X_i) \neq Y_i \bigr] - \mathbb{1} \bigl[f_{\theta'}(X_i) \neq Y_i \bigr] \Bigr|$$

$$\text{and} \quad \overline{m}(\theta, \theta') = \frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} \Bigl| \mathbb{1} \bigl[f_\theta(X_i) \neq Y_i \bigr] - \mathbb{1} \bigl[f_{\theta'}(X_i) \neq Y_i \bigr] \Bigr|,$$

we could prove along the same line of reasoning

THEOREM 2.3. *For any real parameter λ, any $\widetilde{\theta} \in \Theta$, any partially exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda\Bigl[\rho\bigl\{\Psi_{\frac{\lambda}{N}}\bigl[\overline{r}(\cdot) - \overline{r}(\widetilde{\theta}), \overline{m}(\cdot, \widetilde{\theta})\bigr]\bigr\} \\ - \bigl[\rho(r_1) - r_1(\widetilde{\theta})\bigr]\Bigr] - \mathcal{K}(\rho, \pi)\biggr]\biggr\} \leq 1.$$

THEOREM 2.4. *For any real constant γ, for any $\widetilde{\theta} \in \Theta$, for any partially exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \Bigl\{-N\rho\Bigl\{\log\Bigl[1 - \tanh\bigl(\tfrac{\gamma}{N}\bigr)\bigl[\overline{r}(\cdot) - \overline{r}(\widetilde{\theta})\bigr]\Bigr]\Bigr\} \\ - \gamma\bigl[\rho(r_1) - r_1(\widetilde{\theta})\bigr] - N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\rho\bigl[m'(\cdot, \widetilde{\theta})\bigr] - \mathcal{K}(\rho, \pi)\Bigr\}\biggr]\biggr\} \leq 1.$$

This last theorem can be generalized to give

THEOREM 2.5. *For any real constant γ, for any partially exchangeable posterior distributions $\pi^1, \pi^2 : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho_1, \rho_2 \in \mathcal{M}_+^1(\Theta)} \Bigl\{-N\log\Bigl\{1 - \tanh\bigl(\tfrac{\gamma}{N}\bigr)\bigl[\rho_1(\overline{r}) - \rho_2(\overline{r})\bigr]\Bigr\} \\ - \gamma\bigl[\rho_1(r_1) - \rho_2(r_1)\bigr] - N\log\bigl[\cosh\bigl(\tfrac{\gamma}{N}\bigr)\bigr]\rho_1 \otimes \rho_2(m') \\ - \mathcal{K}(\rho_1, \pi^1) - \mathcal{K}(\rho_2, \pi^2)\Bigr\}\biggr]\biggr\} \leq 1.$$

To conclude this section, we see that the basic theorems of transductive PAC-Bayesian classification have exactly the same form as the basic inequalities of inductive classification, Theorems 1.4, 1.25 and 1.26 *with $R(\theta)$ replaced with* $\overline{r}(\theta)$, $r(\theta)$ replaced with $r_1(\theta)$ and $M'(\theta, \widetilde{\theta})$ replaced with $\overline{m}(\theta, \widetilde{\theta})$.

Thus all the results of the first section remain true under the hypotheses of transductive classification, with $R(\theta)$ replaced with $\overline{r}(\theta)$, $r(\theta)$ replaced with $r_1(\theta)$ and $M'(\theta, \widetilde{\theta})$ replaced with $\overline{m}(\theta, \widetilde{\theta})$.

Consequently, in the case when the unlabelled shadow sample is observed, it is possible to improve on Vapnik's bounds to be discussed hereafter by using an explicit partially exchangeable posterior distribution π and resorting

to localized or to relative bounds (in the case at least of unlimited computing resources, which of course may still be unrealistic in many real world situations, and with the caveat, to be recalled in the conclusion of this article, that for small sample sizes and comparatively complex classification models, the improvement may not be so decisive).

Let us notice also that the transductive setting when experimentally available, has the advantage that

$$\overline{d}(\theta,\theta') = \frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} \mathbb{1}\bigl[f_{\theta'}(X_i) \neq f_{\theta}(X_i)\bigr] \\ \geq \overline{m}(\theta,\theta') \geq \overline{r}(\theta) - \overline{r}(\theta'), \qquad \theta,\theta' \in \Theta,$$

is observable in this context, providing an empirical upper bound for the difference $\overline{r}(\widehat{\theta}) - \rho(\overline{r})$ for any non randomized estimator $\widehat{\theta}$ and any posterior distribution ρ, namely

$$\overline{r}(\widehat{\theta}) \leq \rho(\overline{r}) + \rho\bigl[\,\overline{d}(\cdot,\widehat{\theta})\bigr].$$

Thus in the setting of transductive statistical experiments, the PAC-Bayesian framework provides fully empirical bounds for the error rate of non randomized estimators $\widehat{\theta} : \Omega \to \Theta$, even when using a non atomic prior π (or more generally a non atomic partially exchangeable posterior distribution π), when Θ is not a vector space and $\theta \mapsto R(\theta)$ cannot be proved to be convex on the support of some useful posterior distribution ρ.

2.2. Vapnik's bounds for transductive classification.

In this section, we are going to stick to plain unlocalized non relative bounds. As we have already mentioned, (and as it was put forward by Vapnik himself in his seminal works), these bounds are not always superseded by the asymptotically better ones, and deserve all our efforts since they deal in many situations better with small samples.

2.2.1. With a shadow sample of arbitrary size.

The great thing with the transductive setting is that we are manipulating only r_1 and $\overline{r}$ which can take but a finite number of values and therefore are piecewise constant on Θ. To make use of this, let us consider for any value $\theta \in \Theta$ of the parameter the subset $\Delta(\theta) \subset \Theta$ of parameters θ' such that the classification rule $f_{\theta'}$ answers the same on the extended sample $(X_i)_{i=1}^{(k+1)N}$ as f_θ. Namely, let us put for any $\theta \in \Theta$

$$\Delta(\theta) = \bigl\{\theta' \in \Theta; f_{\theta'}(X_i) = f_\theta(X_i), i = 1,\dots,(k+1)N\bigr\}.$$

We see immediately that $\Delta(\theta)$ is an exchangeable parameter subset on which r_1 and r_2 (and therefore also $\overline{r}$) take a constant value. Thus for any $\theta \in \Theta$ we may consider the posterior ρ_θ defined by

$$\frac{d\rho_\theta}{d\pi}(\theta') = \mathbb{1}\bigl[\theta' \in \Delta(\theta)\bigr]\pi\bigl[\Delta(\theta)\bigr]^{-1},$$

and use the fact that $\rho_\theta(r_1) = r_1(\theta)$ and $\rho_\theta(\overline{r}) = \overline{r}(\theta)$, to prove that

Lemma 2.6. *For any partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \rightarrow \mathbb{R}$ such that*

$$\lambda(\omega, \theta') = \lambda(\omega, \theta), \quad \theta \in \Theta, \theta' \in \Delta(\theta), \omega \in \Omega, \tag{2.1}$$

and any partially exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,

$$\Phi_{\frac{\lambda}{N}}\bigl[\overline{r}(\theta)\bigr] + \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda(\theta)} \leq r_1(\theta).$$

We can then remark that for any value of λ independent of ω, the left-hand side of the previous inequality is a partially exchangeable function of $\omega \in \Omega$. Thus this left-hand side is maximized by some partially exchangeable function λ, namely

$$\arg\max_{\lambda} \Phi_{\frac{\lambda}{N}}\bigl[\overline{r}(\theta)\bigr] + \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda}$$

is partially exchangeable as depending only on partially exchangeable quantities. Moreover this choice of $\lambda(\omega, \theta)$ satisfies also condition (2.1) stated in the previous lemma of being constant on $\Delta(\theta)$, proving

Lemma 2.7. *For any partially exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$ and any $\lambda \in \mathbb{R}_+$,*

$$\Phi_{\frac{\lambda}{N}}\bigl[\overline{r}(\theta)\bigr] + \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda} \leq r_1(\theta).$$

Writing $\overline{r} = \frac{r_1 + k r_2}{k+1}$ and rearranging terms we obtain

Theorem 2.8. *For any partially exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$r_2(\theta) \leq \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{1 - \exp\left(-\frac{\lambda}{N} r_1(\theta) + \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{N}\right)}{1 - \exp\bigl(-\frac{\lambda}{N}\bigr)} - \frac{r_1(\theta)}{k}.$$

Let us remind the reader that in the case when we have a set of binary classification rules $\{f_\theta; \theta \in \Theta\}$ whose VC dimension is not greater than h, then we can choose π such that $\pi\bigl[\Delta(\theta)\bigr]$ is independent of θ and not less that $\displaystyle \left(\frac{h}{e(k+1)N}\right)^h$.

Another important case when the complexity term $-\log\bigl\{\pi\bigl[\Delta(\theta)\bigr]\bigr\}$ can easily be controlled is the setting of *compression schemes*, introduced by Littlestone et Warmuth [24]. In this case, we are given for each labelled subsample $(X_i, Y_i)_{i \in J}$, $J \subset \{1, \dots, N\}$, an estimator of the parameter

$$\widehat{\theta}\bigl[(X_i, Y_i)_{i \in J}\bigr] = \widehat{\theta}_J, \quad J \subset \{1, \dots, N\}, |J| \leq h,$$

where

$$\widehat{\theta} : \bigsqcup_{k=1}^{N} \bigl(\mathcal{X} \times \mathcal{Y}\bigr)^k \rightarrow \Theta$$

is an exchangeable function providing estimators for subsamples of arbitrary size. Let us assume that $\widehat{\theta}$ is exchangeable, meaning that for any $k = 1, \dots, N$ and any permutation σ of $\{1, \dots, k\}$

$$\widehat{\theta}\bigl[(x_i, y_i)_{i=1}^k\bigr] = \widehat{\theta}\bigl[(x_{\sigma(i)}, y_{\sigma(i)})_{i=1}^k\bigr], \qquad (x_i, y_i)_{i=1}^k \in \bigl(\mathcal{X} \times \mathcal{Y}\bigr)^k.$$

In this situation, we can introduce the exchangeable subset

$$\Bigl\{\widehat{\theta}_J; J \subset \{1, \dots, (k+1)N\}, |J| \leq h\Bigr\} \subset \Theta,$$

which is seen to contain at most $\displaystyle \sum_{j=0}^{h} \binom{(k+1)N}{j} \leq \left(\frac{e(k+1)N}{h}\right)^h$ classification rules (as will be proved later on in Theorem 3.14 on page 138). Note that we had to extend the range of J to all the subsets of the extended sample, although we will use for estimation only those of the training sample, on which the labels are observed. Thus in this case also we can find a partially exchangeable posterior distribution π such that $\pi\bigl[\Delta(\widehat{\theta}_J)\bigr] \geq \displaystyle \left(\frac{h}{e(k+1)N}\right)^h$. We see that the size of the compression scheme plays the same role in this complexity bound as the VC dimension for VC classes.

In these two cases of binary classification with VC dimension not greater than h and compression schemes depending on a compression set with at most h points, we get a bound of

$$r_2(\theta) \leq \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{1 - \exp\left(- \frac{\lambda}{N} r_1(\theta) - \frac{h \log\left(\frac{e(k+1)N}{h}\right) - \log(\epsilon)}{N}\right)}{1 - \exp\bigl(-\frac{\lambda}{N}\bigr)} - \frac{r_1(\theta)}{k}.$$

Let us make some numerical application: when $N = 1000, h = 10, \epsilon = 0.01$, and $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$, we find that $r_2(\widehat{\theta}) \leq 0.4093$, for k between 15 and 17, and values of λ equal respectively to 965, 968 and 971. For $k = 1$, we find only $r_2(\widehat{\theta}) \leq 0.539$, showing the interest of allowing k to be larger than 1.

2.2.2. When the shadow sample has the same size as the training sample. In the case when $k = 1$, we can improve Theorem 2.2 by taking advantage of the fact that $T_i(\sigma_i)$ can take only 3 values, namely 0, 0.5 and 1. We see thus that $T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\bigl[T_i(\sigma_i)\bigr]$ can take only two values, 0 and $\frac{1}{2} - \Phi_{\frac{\lambda}{N}}(\frac{1}{2})$, because $\Phi_{\frac{\lambda}{N}}(0) = 0$ and $\Phi_{\frac{\lambda}{N}}(1) = 1$. Thus

$$T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\bigl[T_i(\sigma_i)\bigr] = \bigl[1 - |1 - 2T_i(\sigma_i)|\bigr]\bigl[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\bigr].$$

This shows that in the case when $k = 1$,

$$\begin{aligned} \log\Bigl\{T\bigl[\exp(-\lambda r_1)\bigr]\Bigr\} &= -\lambda\overline{r} + \frac{\lambda}{N}\sum_{i=1}^N T_i(\sigma_i) - \Phi_{\frac{\lambda}{N}}\bigl[T_i(\sigma_i)\bigr] \\ &= -\lambda\overline{r} + \frac{\lambda}{N}\sum_{i=1}^N \bigl[1 - |1 - 2T_i(\sigma_i)|\bigr]\bigl[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\bigr] \\ &\leq -\lambda\overline{r} + \lambda\bigl[\tfrac{1}{2} - \Phi_{\frac{\lambda}{N}}(\tfrac{1}{2})\bigr]\bigl[1 - |1 - 2\overline{r}|\bigr]. \end{aligned}$$

Noticing that $\frac{1}{2} - \Phi_{\frac{\lambda}{N}}(\frac{1}{2}) = \frac{N}{\lambda}\log\bigl[\cosh(\frac{\lambda}{2N})\bigr]$, we obtain

THEOREM 2.9. *For any partially exchangeable function $\lambda : \Omega \times \Theta \rightarrow \mathbb{R}_+$, for any partially exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,*

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho\Bigl[\lambda(\overline{r} - r_1) - N\log\bigl[\cosh(\tfrac{\lambda}{2N})\bigr]\bigl(1 - |1 - 2\overline{r}|\bigr)\Bigr] - \mathcal{K}(\rho, \pi)\biggr]\biggr\} \leq 1.$$

As a consequence, reasonning as previously, we deduce

THEOREM 2.10. *In the case when $k = 1$, for any partially exchangeable posterior distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$ and any $\lambda \in \mathbb{R}_+$,*

$$\overline{r}(\theta) - \tfrac{N}{\lambda} \log\bigl[\cosh(\tfrac{\lambda}{2N})\bigr]\bigl(1 - |1 - 2\overline{r}(\theta)|\bigr) + \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda} \leq r_1(\theta);$$

and consequently for any $\theta \in \Theta$,

$$r_2(\theta) \leq 2 \inf_{\lambda \in \mathbb{R}_+} \frac{r_1(\theta) - \dfrac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda}}{1 - \frac{2N}{\lambda}\log\bigl[\cosh(\frac{\lambda}{2N})\bigr]} - r_1(\theta).$$

In the case of binary classification using a VC class of VC dimension not greater than h, we can choose π such that $-\log\bigl\{\pi\bigl[\Delta(\theta)\bigr]\bigr\} \leq h\log(\frac{2eN}{h})$ and obtain the following numerical illustration of this theorem : for $N = 1000$, $h = 10$, $\epsilon = 0.01$ and $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$, we find an upper bound $r_2(\widehat{\theta}) \leq 0.5033$, which improves on Theorem 2.8 but still is not under the significance level $\frac{1}{2}$ (achieved by blind random classification). This indicates that considering shadow samples of arbitrary sizes brings in some noisy situations a significant improvement on bounds obtained with a shadow sample of the same size as the training sample.

2.2.3. When moreover the distribution of the augmented sample is exchangeable. In the case when $k = 1$ and $\mathbb{P}$ is exchangeable meaning that for any bounded measurable function $h : \Omega \to \mathbb{R}$ and any permutation $s \in \mathfrak{S}\bigl(\{1, \dots, 2N\}\bigr)$ $\mathbb{P}\bigl[h(\omega \circ s)\bigr] = \mathbb{P}\bigl[h(\omega)\bigr]$, then we can still improve the bound as follows. Let

$$T'(h) = \frac{1}{N!} \sum_{s \in \mathfrak{S}\left(\{N+1, \dots, 2N\}\right)} h(\omega \circ s).$$

Then we can write

$$1 - |1 - 2T_i(\sigma_i)| = (\sigma_i - \sigma_{i+N})^2 = \sigma_i + \sigma_{i+N} - 2\sigma_i\sigma_{i+N}.$$

Using this identity, we get for any exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$,

$$T\biggl\{\exp\biggl[\lambda(\overline{r} - r_1) - \log\bigl[\cosh(\tfrac{\lambda}{2N})\bigr] \sum_{i=1}^N \bigl(\sigma_i + \sigma_{i+N} - 2\sigma_i\sigma_{i+N}\bigr)\biggr]\biggr\} \leq 1.$$

Let us put

$$A(\lambda) = \tfrac{2N}{\lambda} \log \bigl[\cosh (\tfrac{\lambda}{2N}) \bigr], \tag{2.2}$$

$$v(\theta) = \frac{1}{2N} \sum_{i=1}^{N} (\sigma_i + \sigma_{i+N} - 2 \sigma_i \sigma_{i+N}). \tag{2.3}$$

With these notations

$$T \Bigl\{ \exp \bigl\{ \lambda \bigl[\overline{r} - r_1 - A(\lambda) v \bigr] \bigr\} \Bigr\} \leq 1.$$

Let notice now that

$$T' \bigl[v(\theta) \bigr] = \overline{r}(\theta) - r_1(\theta) r_2(\theta).$$

Let $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ be any given exchangeable posterior distribution. Using the exchangeability of $\mathbb{P}$ and π and the exchangeability of the exponential function, we get

$$\begin{aligned}
\mathbb{P} \Bigl\{ \pi \Bigl[\exp \bigl\{ \lambda \bigl[\overline{r} - r_1 - A(\overline{r} - r_1 r_2) \bigr] \bigr\} \Bigr] \Bigr\} &= \mathbb{P} \Bigl\{ \pi \Bigl[\exp \bigl\{ \lambda \bigl[\overline{r} - r_1 - A T'(v) \bigr] \bigr\} \Bigr] \Bigr\} \\
\leq \mathbb{P} \Bigl\{ \pi \Bigl[T' \exp \bigl\{ \lambda \bigl[\overline{r} - r_1 - A v \bigr] \bigr\} \Bigr] \Bigr\} &= \mathbb{P} \Bigl\{ T' \pi \Bigl[\exp \bigl\{ \lambda \bigl[\overline{r} - r_1 - A v \bigr] \bigr\} \Bigr] \Bigr\} \\
= \mathbb{P} \Bigl\{ \pi \Bigl[\exp \bigl\{ \lambda \bigl[\overline{r} - r_1 - A v \bigr] \bigr\} \Bigr] \Bigr\} &= \mathbb{P} \Bigl\{ T \pi \Bigl[\exp \bigl\{ \lambda \bigl[\overline{r} - r_1 - A v \bigr] \bigr\} \Bigr] \Bigr\} \\
= \mathbb{P} \Bigl\{ \pi \Bigl[T \exp \bigl\{ \lambda \bigl[\overline{r} - r_1 - A v \bigr] \bigr\} \Bigr] \Bigr\} &\leq 1.
\end{aligned}$$

We are thus ready to state

Theorem 2.11. *In the case when $k = 1$, for any exchangeable probability distribution $\mathbb{P}$, for any exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, for any exchangeable function $\lambda : \Omega \times \Theta \rightarrow \mathbb{R}_+$,*

$$\mathbb{P} \Biggl\{ \exp \Biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho \Bigl\{ \lambda \bigl[\overline{r} - r_1 - A(\lambda) (\overline{r} - r_1 r_2) \bigr] \Bigr\} - \mathcal{K}(\rho, \pi) \Biggr] \Biggr\} \leq 1,$$

where $A(\lambda)$ is defined by equation (2.2) *above.*

We then deduce as previously

Corollary 2.12. *For any exchangeable posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$, for any exchangeable probability measure $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, for any measurable exchangeable function $\lambda : \Omega \times \Theta \rightarrow \mathbb{R}_+$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$\overline{r}(\theta) \leq r_1(\theta) + A(\lambda) \bigl[\overline{r}(\theta) - r_1(\theta) r_2(\theta) \bigr] - \frac{\log \bigl\{ \epsilon \pi \bigl[\Delta(\theta) \bigr] \bigr\}}{\lambda},$$

where $A(\lambda)$ is defined by equation (2.2) *on page 108.*

In order to deduce an empirical bound from this theorem, we have to make some choice for $\lambda(\omega, \theta)$. Fortunately, it is easy to show that the bound indeed holds uniformly in λ. This is the case because the inequality can be rewritten as a function of only one non exchangeable quantity, namely $r_1(\theta)$. Indeed, since $r_2 = 2\overline{r} - r_1$, we see that the inequality can be written as

$$\overline{r}(\theta) \leq r_1(\theta) + A(\lambda)\bigl[\overline{r}(\theta) - 2\overline{r}(\theta) r_1(\theta) + r_1(\theta)^2\bigr] - \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]}{\lambda}.$$

It can be solved in $r_1(\theta)$, to get

$$r_1(\theta) \geq f\Bigl(\lambda, \overline{r}(\theta), -\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}\Bigr),$$

where namely

$$f(\lambda, \overline{r}, d) = \bigl[2A(\lambda)\bigr]^{-1}\Biggl\{2\overline{r}A(\lambda) - 1 + \sqrt{\bigl[1 - 2\overline{r}A(\lambda)\bigr]^2 + 4A(\lambda)\Bigl\{\overline{r}\bigl[1 - A(\lambda)\bigr] - \tfrac{d}{\lambda}\Bigr\}}\Biggr\}.$$

Thus we can find some exchangeable function $\lambda(\omega, \theta)$, such that

$$f\Bigl(\lambda(\omega,\theta), \overline{r}(\theta), -\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}\Bigr) = \sup_{\beta \in \mathbb{R}_+} f\Bigl(\beta, \overline{r}(\theta), -\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}\Bigr).$$

Applying Corollary 2.12 to that choice of λ, we see that

Theorem 2.13. *For any exchangeable probability measure $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, for any exchangeable posterior probability distribution $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$, for any $\lambda \in \mathbb{R}_+$,*

$$\overline{r}(\theta) \leq r_1(\theta) + A(\lambda)\bigl[\overline{r}(\theta) - r_1(\theta) r_2(\theta)\bigr] - \frac{\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda},$$

where $A(\lambda)$ is defined by equation (2.2) *on page 108.*

Solving the previous inequality in $r_2(\theta)$, we get

Corollary 2.14. *Under the same assumptions as in the previous theorem, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$r_2(\theta) \leq \inf_{\lambda \in \mathbb{R}_+} \frac{r_1(\theta)\Bigl\{1 + \frac{2N}{\lambda}\log\bigl[\cosh(\frac{\lambda}{2N})\bigr]\Bigr\} - \dfrac{2\log\bigl\{\epsilon\pi\bigl[\Delta(\theta)\bigr]\bigr\}}{\lambda}}{1 - \frac{2N}{\lambda}\log\bigl[\cosh(\frac{\lambda}{2N})\bigr]\bigl[1 - 2r_1(\theta)\bigr]}.$$

Applying this to our usual numerical example of a binary classification model with VC dimension not greater than $h = 10$, when $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 10$ and $\epsilon = 0.01$, we obtain that $r_2(\widehat{\theta}) \leq 0.4450$.

2.3. Vapnik's bounds for inductive classification.

2.3.1. Arbitrary shadow sample size. We assume in this section that

$$\mathbb{P} = \biggl(\bigotimes_{i=1}^N P_i \biggr)^{\otimes \infty} \in \mathcal{M}_+^1 \Bigl\{ \bigl[(\mathcal{X} \times \mathcal{Y})^N \bigr]^{\mathbb{N}} \Bigr\},$$

where $P_i \in \mathcal{M}_+^1\bigl(\mathcal{X} \times \mathcal{Y}\bigr)$: we consider an infinite i.i.d. sequence of independent *not* identically distributed samples of size N, the first one only being observed. The shadow samples will only appear in the proofs. The aim of this section is to prove better Vapnik's bounds, generalizing them in the same time to the independent non i.i.d. setting, which to our knowledge had not been done before.

Let us introduce the notation $\mathbb{P}'\bigl[h(\omega)\bigr] = \mathbb{P}\bigl[h(\omega) \,|\, (X_i, Y_i)_{i=1}^N\bigr]$, where h may be any suitable (e.g. bounded) random variable, let us also put $\Omega = \bigl[(\mathcal{X} \times \mathcal{Y})^N\bigr]^{\mathbb{N}}$.

Definition 2.2. For any subset $A \subset \mathbb{N}$ of integers, let $\mathfrak{C}(A)$ be the set of circular permutations of the totally ordered set A, extended to a permutation of $\mathbb{N}$ by taking it to be the identity on the complement $\mathbb{N} \setminus A$ of A. We will say that a random function $h : \Omega \rightarrow \mathbb{R}$ is k-partially exchangeable if

$$h(\omega \circ s) = h(\omega), \quad s \in \mathfrak{C}\bigl(\{i + jN \,;\, j = 0, \dots, k\}\bigr), i = 1, \dots, N.$$

In the same way, we will say that a posterior distribution $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ is k-partially exchangeable if

$$\pi(\omega \circ s) = \pi(\omega) \in \mathcal{M}_+^1(\Theta), \quad s \in \mathfrak{C}\bigl(\{i + jN \,;\, j = 0, \dots, k\}\bigr), i = 1, \dots, N.$$

Note that $\mathbb{P}$ itself is k-partially exchangeable for any k in the sense that for any bounded measurable function $h : \Omega \rightarrow \mathbb{R}$

$$\mathbb{P}\bigl[h(\omega \circ s)\bigr] = \mathbb{P}\bigl[h(\omega)\bigr], \quad s \in \mathfrak{C}\bigl(\{i + jN \,;\, j = 0, \dots, k\}\bigr), i = 1, \dots, N.$$

Let $\Delta_k(\theta) = \Bigl\{ \theta' \in \Theta \,;\, \bigl[f_{\theta'}(X_i) \bigr]_{i=1}^{(k+1)N} = \bigl[f_\theta(X_i) \bigr]_{i=1}^{(k+1)N} \Bigr\}$, $\theta \in \Theta, k \in \mathbb{N}^*$, and let also $\overline{r}_k(\theta) = \dfrac{1}{(k+1)N} \displaystyle\sum_{i=1}^{(k+1)N} \mathbb{1}\bigl[f_\theta(X_i) \neq Y_i \bigr]$. Theorem 2.2 shows

that for any positive real parameter λ and any k-partially exchangeable posterior distribution $\pi_k : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\biggl\{\exp\biggl[\sup_{\theta\in\Theta}\lambda\bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1\bigr] + \log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\biggr]\biggr\} \leq \epsilon.$$

Using the general fact that

$$\mathbb{P}\bigl[\exp(h)\bigr] = \mathbb{P}\Bigl\{\mathbb{P}'\bigl[\exp(h)\bigr]\Bigr\} \geq \mathbb{P}\Bigl\{\exp\bigl[\mathbb{P}'(h)\bigr]\Bigr\},$$

and the fact that the expectation of a supremum is larger than the supremum of an expectation, we see that with $\mathbb{P}$ probability at most $1-\epsilon$, for any $\theta \in \Theta$,

$$\mathbb{P}'\Bigl\{\Phi_{\frac{\lambda}{N}}\bigl[\overline{r}_k(\theta)\bigr]\Bigr\} \leq r_1(\theta) - \frac{\mathbb{P}'\Bigl\{\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\Bigr\}}{\lambda}.$$

Let us put for short

$$\begin{aligned}
\bar{d}_k(\theta) &= -\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\},\\
d_k'(\theta) &= -\mathbb{P}'\Bigl\{\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\Bigr\},\\
d_k(\theta) &= -\mathbb{P}\Bigl\{\log\bigl\{\epsilon\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\Bigr\}.
\end{aligned}$$

We can use the convexity of $\Phi_{\frac{\lambda}{N}}$ and the fact that $\mathbb{P}'(\overline{r}_k) = \frac{r_1+kR}{k+1}$, to see that

$$\mathbb{P}'\Bigl\{\Phi_{\frac{\lambda}{N}}\bigl[\overline{r}_k(\theta)\bigr]\Bigr\} \geq \Phi_{\frac{\lambda}{N}}\left[\frac{r_1(\theta)+kR(\theta)}{k+1}\right].$$

We have proved

THEOREM 2.15. *Using the above hypotheses and notations, for any sequence* $\pi_k : \Omega \to \mathcal{M}_+^1(\Theta)$, *where* π_k *is a* k*-partially exchangeable posterior distribution, for any positive real constant* λ, *any positive integer* k, *with* $\mathbb{P}$ *probability at least* $1-\epsilon$, *for any* $\theta \in \Theta$,

$$\Phi_{\frac{\lambda}{N}}\left[\frac{r_1(\theta)+kR(\theta)}{k+1}\right] \leq r_1(\theta) + \frac{d_k'(\theta)}{\lambda}.$$

We can make as we did with Theorem 1.10 on page 20 the result of this theorem uniform in $\lambda \in \{\alpha^j\,;\, j \in \mathbb{N}^*\}$ and $k \in \mathbb{N}^*$ (considering on k the prior $\frac{1}{k(k+1)}$ and on j the prior $\frac{1}{j(j+1)}$), and obtain

Theorem 2.16. *For any real parameter $\alpha > 1$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{k \in \mathbb{N}^*, j \in \mathbb{N}^*} \frac{1 - \exp\Bigl\{ - \frac{\alpha^j}{N} r_1(\theta) - \frac{1}{N} \Bigl\{ d'_k(\theta) + \log \bigl[k(k+1) j(j+1) \bigr] \Bigr\} \Bigr\}}{\frac{k}{k+1} \Bigl[1 - \exp \Bigl(- \frac{\alpha^j}{N} \Bigr) \Bigr]} - \frac{r_1(\theta)}{k}.$$

Note that as a special case we can choose π_k such that $\log\bigl\{\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}$ is independent of θ and equal to $\log(\mathfrak{N}_k)$, where $\mathfrak{N}_k = \bigl|\bigl\{\bigl[f_\theta(X_i)\bigr]_{i=1}^{(k+1)N}\,;\, \theta \in \Theta\bigr\}\bigr|$ is the size of the trace of the classification model on the extended sample of size $(k+1)N$. With this choice, we obtain a bound involving a new flavour of conditional Vapnik's entropy, namely

$$d'_k(\theta) = \mathbb{P}\bigl[\log(\mathfrak{N}_k) \,|\, (Z_i)_{i=1}^N\bigr] - \log(\epsilon).$$

In the case of binary classification using a VC class of VC dimension not greater than $h = 10$, when $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$ and $\epsilon = 0.01$, choosing $\alpha = 1.1$, we obtain $R(\widehat{\theta}) \leq 0.4271$ (for an optimal value of $\lambda = 1071.8$, and an optimal value of $k = 16$).

2.3.2. A better minimization with respect to the exponential parameter. If we are not pleased with the fact of optimizing λ on a discrete subset of the real line, we can use a slightly different approach. From Theorem 2.2, we see that for any positive integer k, for any k-partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \rightarrow \mathbb{R}_+$ satisfying equation (2.1) on page 104 (with $\Delta(\theta)$ replaced with $\Delta_k(\theta)$), for any $\epsilon \in)0,1)$ and $\eta \in)0,1)$,

$$\mathbb{P}\biggl\{\mathbb{P}'\biggl[\exp\Bigl[\sup_\theta \lambda\bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1\bigr] + \log\bigl\{\epsilon\eta\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\Bigr]\biggr]\biggr\} \leq \epsilon\eta,$$

therefore with $\mathbb{P}$ probability at least $1-\epsilon$,

$$\mathbb{P}'\biggl\{\exp\Bigl[\sup_\theta \lambda\bigl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1\bigr] + \log\bigl\{\epsilon\eta\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}\Bigr]\biggr\} \leq \eta,$$

and consequently, with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\theta \in \Theta$,

$$\Phi_{\frac{\lambda}{N}}(\overline{r}_k) + \frac{\log\bigl\{\epsilon\eta\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}}{\lambda} \leq r_1.$$

Now we are entitled to choose

$$\lambda(\omega,\theta) \in \arg\max_{\lambda'\in\mathbb{R}_+} \Phi_{\frac{\lambda'}{N}}(\overline{r}_k) + \frac{\log\bigl\{\epsilon\eta\pi_k\bigl[\Delta_k(\theta)\bigr]\bigr\}}{\lambda'}.$$

This shows that with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\theta \in \Theta$,

$$\sup_{\lambda\in\mathbb{R}_+} \Phi_{\frac{\lambda}{N}}(\overline{r}_k) - \frac{\bar{d}_k(\theta) - \log(\eta)}{\lambda} \leq r_1,$$

which can also be written

$$\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda} \leq -\frac{\log(\eta)}{\lambda}, \quad \lambda \in \mathbb{R}_+.$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, any $\lambda \in \mathbb{R}_+$,

$$\mathbb{P}'\biggl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda}\biggr] \leq -\frac{\log(\eta)}{\lambda} + \biggl[1 - r_1 + \frac{\log(\eta)}{\lambda}\biggr]\eta.$$

On the other hand, $\Phi_{\frac{\lambda}{N}}$ being a convex function,

$$\begin{aligned}\mathbb{P}'\biggl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k(\theta)}{\lambda}\biggr] &\geq \Phi_{\frac{\lambda}{N}}\bigl[\mathbb{P}'(\overline{r}_k)\bigr] - r_1 - \frac{d'_k}{\lambda} \\ &= \Phi_{\frac{\lambda}{N}}\biggl(\frac{kR + r_1}{k+1}\biggr) - r_1 - \frac{d'_k}{\lambda}.\end{aligned}$$

Thus with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,

$$\frac{kR + r_1}{k+1} \leq \inf_{\lambda\in\mathbb{R}_+} \Phi^{-1}_{\frac{\lambda}{N}}\biggl[r_1(1-\eta) + \eta + \frac{d'_k - \log(\eta)(1-\eta)}{\lambda}\biggr].$$

We can generalize this approach by considering a finite decreasing sequence $\eta_0 = 1 > \eta_1 > \eta_2 > \cdots > \eta_J > \eta_{J+1} = 0$, and the corresponding sequence of levels

$$\begin{aligned}L_j &= -\frac{\log(\eta_j)}{\lambda}, 0 \leq j \leq J, \\ L_{J+1} &= 1 - r_1 - \frac{\log(J) - \log(\epsilon)}{\lambda}.\end{aligned}$$

Taking a union bound in j, we see that with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, for any $\lambda \in \mathbb{R}_+$,

$$\mathbb{P}'\biggl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda} \geq L_j\biggr] \leq \eta_j, \quad j = 0, \dots, J+1,$$

and consequently

$$\begin{aligned}
&\mathbb{P}'\biggl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda}\biggr] \\
&\quad \leq \int_0^{L_{J+1}} \mathbb{P}'\biggl[\Phi_{\frac{\lambda}{N}}(\overline{r}_k) - r_1 - \frac{\bar{d}_k + \log(J)}{\lambda} \geq \alpha\biggr] d\alpha \quad \leq \sum_{j=1}^{J+1} \eta_{j-1}(L_j - L_{j-1}) \\
&\quad = \eta_J\biggl[1 - r_1 - \frac{\log(J) - \log(\epsilon) - \log(\eta_J)}{\lambda}\biggr] - \frac{\log(\eta_1)}{\lambda} + \sum_{j=1}^{J-1} \frac{\eta_j}{\lambda} \log\biggl(\frac{\eta_j}{\eta_{j+1}}\biggr).
\end{aligned}$$

Let us put

$$\begin{aligned}
d''_k\bigl[\theta, (\eta_j)_{j=1}^J\bigr] = d'_k(\theta) + \log(J) - \log(\eta_1) \\
+ \sum_{j=1}^{J-1} \eta_j \log\biggl(\frac{\eta_j}{\eta_{j+1}}\biggr) + \log\Bigl(\frac{\epsilon \eta_J}{J}\Bigr)\eta_J.
\end{aligned}$$

We have proved that for any decreasing sequence $(\eta_j)_{j=1}^J$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,

$$\frac{kR + r_1}{k+1} \leq \inf_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1}\biggl[r_1(1 - \eta_J) + \eta_J + \frac{d''_k\bigl[\theta, (\eta_j)_{j=1}^J\bigr]}{\lambda}\biggr].$$

Remark 2.1. We can for instance choose $J = 2$, $\eta_2 = \frac{1}{10N}$, $\eta_1 = \frac{1}{\log(10N)}$, resulting in

$$d''_k = d'_k + \log(2) + \log\log(10N) + 1 - \frac{\log\log(10N)}{\log(10N)} - \frac{\log\bigl(\frac{20N}{\epsilon}\bigr)}{10N}.$$

In the case when $N = 1000$ and for any $\epsilon \in (0,1)$, we get $d''_k \leq d'_k + 3.7$, in the case when $N = 10^6$, we get $d''_k \leq d'_k + 4.4$, and in the case $N = 10^9$, we get $d''_k \leq d'_k + 4.7$.

Therefore, for any practical purpose we could take $d''_k = d'_k + 4.7$ and $\eta_J = \frac{1}{10N}$ in the above inequality.

Taking moreover a weighted union bound in k, we get

Theorem 2.17. *For any $\epsilon \in]0,1)$, any sequence $1 > \eta_1 > \cdots > \eta_J > 0$, any sequence $\pi_k : \Omega \to \mathcal{M}_+^1(\Theta)$, where π_k is a k-partially exchangeable posterior distribution, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{k \in \mathbb{N}^*} \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1} \biggl[r_1(\theta) + \eta_J \bigl[1 - r_1(\theta) \bigr] \\ + \frac{d_k''\bigl[\theta, (\eta_j)_{j=1}^J\bigr] + \log\bigl[k(k+1)\bigr]}{\lambda} \biggr] - \frac{r_1(\theta)}{k}.$$

COROLLARY 2.18. *For any* $\epsilon \in)0,1)$, *for any* $N \leq 10^9$, *with* $\mathbb{P}$ *probability at least* $1-\epsilon$, *for any* $\theta \in \Theta$,

$$R(\theta) \leq \inf_{k \in \mathbb{N}^*} \inf_{\lambda \in \mathbb{R}_+} \frac{k+1}{k} \bigl[1 - \exp(-\tfrac{\lambda}{N}) \bigr]^{-1} \biggl\{ 1 - \exp \biggl[-\tfrac{\lambda}{N} \bigl[r_1(\theta) + \tfrac{1}{10N} \bigr] \\ - \frac{\mathbb{P}'\bigl[\log(\mathfrak{N}_k) \,|\, (Z_i)_{i=1}^N\bigr] - \log(\epsilon) + \log\bigl[k(k+1)\bigr] + 4.7}{N} \biggr] \biggr\} - \frac{r_1(\theta)}{k}.$$

Let us end this section with a numerical example: in the case of binary classification with a VC class of dimension not greater than 10, when $N = 1000$, $\inf_\Theta r_1 = r_1(\widehat{\theta}) = 0.2$ and $\epsilon = 0.01$, we get a bound $R(\widehat{\theta}) \leq 0.4211$ (for optimal values of $k = 15$ and of $\lambda = 1010$).

2.3.3. Equal shadow and training sample sizes. In the case when $k = 1$, we can use Theorem 2.10, and replace $\Phi_{\frac{\lambda}{N}}^{-1}(q)$ with $\bigl\{ 1 - \frac{2N}{\lambda} \log\bigl[\cosh(\frac{\lambda}{2N})\bigr] \bigr\}^{-1} q$, resulting in

THEOREM 2.19. *For any* $\epsilon \in)0,1)$, *any* $N \leq 10^9$, *any 1-partially exchangeable posterior distribution* $\pi_1 : \Omega \to \mathcal{M}_+^1(\Theta)$, *with* $\mathbb{P}$ *probability at least* $1-\epsilon$, *for any* $\theta \in \Theta$,

$$R(\theta) \leq \inf_{\lambda \in \mathbb{R}_+} \frac{\Bigl\{ 1 + \frac{2N}{\lambda} \log\bigl[\cosh(\frac{\lambda}{2N})\bigr] \Bigr\} r_1(\theta) + \dfrac{1}{5N} + 2 \dfrac{d_1'(\theta) + 4.7}{\lambda}}{1 - \frac{2N}{\lambda} \log\bigl[\cosh(\frac{\lambda}{2N})\bigr]}.$$

2.3.4. Improvement on the equal sample size bound in the i.i.d. case. Eventually, in the case when $\mathbb{P}$ is i.i.d., meaning that all the P_i are equal, we can improve the previous bound. For any partially exchangeable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$, we saw in the discussion preceding Theorem 2.11 on page 108 that

$$T \Bigl[\exp\bigl[\lambda(\overline{r}_k - r_1) - A(\lambda) v \bigr] \Bigr] \leq 1,$$

with the notations introduced therein. Thus for any partially exchangeable positive real measurable function $\lambda : \Omega \times \Theta \to \mathbb{R}_+$ satisfying equation (2.1)

on page 104, any 1-partially exchangeable posterior distribution $\pi_1 : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$,

$$\mathbb{P}\biggl\{\exp\Bigl[\sup_{\theta\in\Theta}\lambda\bigl[\overline{r}_k(\theta)-r_1(\theta)-A(\lambda)v(\theta)\bigr]+\log\bigl[\epsilon\pi_1\bigl[\Delta(\theta)\bigr]\bigr]\Bigr]\biggr\}\leq 1.$$

Therefore with $\mathbb{P}$ probability at least $1-\epsilon$, with $\mathbb{P}'$ probability $1-\eta$,

$$\overline{r}_k(\theta)\leq r_1(\theta)+A(\lambda)v(\theta)+\frac{1}{\lambda}\bigl[\bar{d}_1(\theta)-\log(\eta)\bigr]$$

We can then choose $\lambda(\omega,\theta)\in\arg\min_{\lambda'\in\mathbb{R}_+} A(\lambda')v(\theta)+\dfrac{\bar{d}_1(\theta)-\log(\eta)\bigr]}{\lambda'}$, which satisfies the required conditions, to show that with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$, with $\mathbb{P}'$ probability at least $1-\eta$, for any $\lambda\in\mathbb{R}_+$,

$$\overline{r}_k(\theta)\leq r_1(\theta)+A(\lambda)v(\theta)+\frac{\bar{d}_1(\theta)-\log(\eta)}{\lambda}.$$

We can then take a union bound on a decreasing sequence of J values $\eta_1\geq\cdots\geq\eta_J$ of η. Weakening a little the order of quantifiers, we then obtain the following statement: with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta\in\Theta$, for any $\lambda\in\mathbb{R}_+$, for any $j=1,\dots,J$

$$\mathbb{P}'\left[\overline{r}_k(\theta)-r_1(\theta)-A(\lambda)v(\theta)-\frac{\bar{d}_1(\theta)+\log(J)}{\lambda}\geq-\frac{\log(\eta_j)}{\lambda}\right]\leq\eta_j.$$

Consequently for any $\lambda\in\mathbb{R}_+$,

$$\begin{aligned}\mathbb{P}'\left[\overline{r}_k(\theta)-r_1(\theta)-A(\lambda)v(\theta)-\frac{\bar{d}_1(\theta)+\log(J)}{\lambda}\right]&\\ \leq-\frac{\log(\eta_1)}{\lambda}+\eta_J\left[1-r_1(\theta)-\frac{\log(J)-\log(\epsilon)-\log(\eta_J)}{\lambda}\right]&\\ +\sum_{j=1}^{J-1}\frac{\eta_j}{\lambda}\log\left(\frac{\eta_j}{\eta_{j+1}}\right)&.\end{aligned}$$

Moreover $\mathbb{P}'\bigl[v(\theta)\bigr]=\frac{r_1+R}{2}-r_1R$, (this is where we need equidistribution) thus proving that

$$\frac{R-r_1}{2}\leq\frac{A(\lambda)}{2}\Bigl[R+r_1-2r_1R\Bigr]+\frac{d_1''\bigl[\theta,(\eta_j)_{j=1}^J\bigr]}{\lambda}+\eta_J\bigl[1-r_1(\theta)\bigr].$$

Keeping track of quantifiers, we obtain

Theorem 2.20. *For any decreasing sequence* $(\eta_j)_{j=1}^J$*, any* $\epsilon \in (0,1)$*, any 1-partially exchangeable posterior distribution* $\pi : \Omega \to \mathcal{M}_+^1(\Theta)$*, with* $\mathbb{P}$ *probability at least* $1-\epsilon$*, for any* $\theta \in \Theta$*,*

$$R(\theta) \leq \inf_{\lambda \in \mathbb{R}_+} \frac{\Bigl\{1 + \frac{2N}{\lambda} \log\bigl[\cosh(\frac{\lambda}{2N})\bigr]\Bigr\} r_1(\theta) + \dfrac{2d_1''\bigl[\theta, (\eta_j)_{j=1}^J\bigr]}{\lambda} + 2\eta_J\bigl[1 - r_1(\theta)\bigr]}{1 - \frac{2N}{\lambda}\log\bigl[\cosh(\frac{\lambda}{2N})\bigr]\bigl[1 - 2r_1(\theta)\bigr]}.$$

2.4. Gaussian approximation in Vapnik's bounds. To obtain formulas which could be easily compared with original Vapnik's bounds, we may replace $p - \Phi_a(p)$ with a Gaussian upper bound:

Lemma 2.21. *For any* $p \in (0, \frac{1}{2})$*, any* $a \in \mathbb{R}_+$*,*

$$p - \Phi_a(p) \leq \frac{a}{2} p(1-p).$$

For any $p \in (\frac{1}{2}, 1)$*,*

$$p - \Phi_a(p) \leq \frac{a}{8}.$$

Proof. Let us notice that for any $p \in (0,1)$,

$$\begin{aligned}
\frac{\partial}{\partial a}\bigl[-a\Phi_a(p)\bigr] &= -\frac{p\exp(-a)}{1-p+p\exp(-a)},\\
\frac{\partial^2}{\partial^2 a}\bigl[-a\Phi_a(p)\bigr] &= \frac{p\exp(-a)}{1-p+p\exp(-a)}\left(1 - \frac{p\exp(-a)}{1-p+p\exp(-a)}\right)\\
&\leq \begin{cases} p(1-p) & p \in (0,\frac{1}{2}),\\ \frac{1}{4} & p \in (\frac{1}{2},1).\end{cases}
\end{aligned}$$

Thus taking a Taylor expansion of order one with integral remainder :

$$-a\Phi(a) \leq \begin{cases} -ap + \displaystyle\int_0^a p(1-p)(a-b)db \\ \qquad\qquad = -ap + \dfrac{a^2}{2}p(1-p), & p \in (0,\frac{1}{2}),\\ -ap + \displaystyle\int_0^a \frac{1}{4}(a-b)db = -ap + \frac{a^2}{8}, & p \in (\frac{1}{2},1).\end{cases}$$

This ends the proof of our lemma. □

Lemma 2.22. *Let us consider the bound*

$$B(q,d) = \left(1 + \frac{2d}{N}\right)^{-1} \left[q + \frac{d}{N} + \sqrt{\frac{2dq(1-q)}{N} + \frac{d^2}{N^2}} \right], \quad q \in \mathbb{R}_+, d \in \mathbb{R}_+.$$

Let us also put

$$\bar{B}(q,d) = \begin{cases} B(q,d) & B(q,d) \leq \frac{1}{2}, \\ q + \sqrt{\frac{d}{2N}} & \textit{otherwise.} \end{cases}$$

For any positive real parameters q and d

$$\inf_{\lambda \in \mathbb{R}_+} \Phi_{\frac{\lambda}{N}}^{-1}\left(q + \frac{d}{\lambda}\right) \leq \bar{B}(q,d).$$

Proof. Let $p = \inf_{\lambda} \Phi_{\frac{\lambda}{N}}^{-1}\left(q + \frac{d}{\lambda}\right)$. For any $\lambda \in \mathbb{R}_+$,

$$p - \frac{\lambda}{2N}(p \wedge \tfrac{1}{2})\bigl[1 - (p \wedge \tfrac{1}{2})\bigr] \leq \Phi_{\frac{\lambda}{N}}(p) \leq q + \frac{d}{\lambda}.$$

Thus

$$\begin{aligned} p \leq q + \inf_{\lambda \in \mathbb{R}_+} \frac{\lambda}{2N}(p \wedge \tfrac{1}{2})\bigl[1 - (p \wedge \tfrac{1}{2})\bigr] + \frac{d}{\lambda} \\ = q + \sqrt{\frac{2d(p \wedge \frac{1}{2})\bigl[1 - (p \wedge \frac{1}{2})\bigr]}{N}} \leq q + \sqrt{\frac{d}{2N}}. \end{aligned}$$

Then let us remark that $B(q,d) = \sup \left\{ p' \in \mathbb{R}_+ \,;\, p' \leq q + \sqrt{\frac{2dp'(1-p')}{N}} \right\}$.
If moreover $\frac{1}{2} \geq B(q,d)$, then according to this remark $\frac{1}{2} \geq q + \sqrt{\frac{d}{2N}} \geq p$. Therefore $p \leq \frac{1}{2}$, and consequently $p \leq q + \sqrt{\frac{2dp(1-p)}{N}}$, implying that $p \leq B(q,d)$. $\square$

2.4.1. Arbitrary shadow sample size. This lemma combined with Corollary 2.18 on page 115 implies

Corollary 2.23. *For any $\epsilon \in)0,1)$, any integer $N \leq 10^9$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq \inf_{k \in \mathbb{N}^*} \frac{k+1}{k} \left\{ \bar{B}\Bigl[r_1(\theta) + \frac{1}{10N}, d_k'(\theta) + \log\bigl[k(k+1)\bigr] + 4.7 \Bigr] \right\} - \frac{r_1(\theta)}{k}.$$

2.4.2. Equal sample sizes in the i.i.d. case. To make a link with Vapnik's result, it is useful to work out the Gaussian approximation to Theorem 2.20 on page 117. Indeed, using the upper bound $A(\lambda) \leq \frac{\lambda}{4N}$, where $A(\lambda)$ is defined by equation (2.2) on page 108, we get with $\mathbb{P}$ probability at least $1-\epsilon$

$$R - r_1 - 2\eta_J \leq \inf_{\lambda \in \mathbb{R}_+} \frac{\lambda}{4N} \bigl[R + r_1 - 2 r_1 R \bigr] + \frac{2 d_1''}{\lambda} = \sqrt{\frac{2 d_1''(R + r_1 - 2 r_1 R)}{N}},$$

which can be solved in R to obtain

Corollary 2.24. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$,*

$$R(\theta) \leq r_1(\theta) + \frac{d_1''(\theta)}{N} \bigl[1 - 2 r_1(\theta) \bigr] + 2\eta_J$$
$$+ \sqrt{\frac{4 d_1''(\theta) \bigl[1 - r_1(\theta) \bigr] r_1(\theta)}{N} + \frac{d_1''(\theta)^2}{N^2} \bigl[1 - 2 r_1(\theta) \bigr]^2 + \frac{4 d_1''(\theta)}{N} \bigl[1 - 2 r_1(\theta) \bigr] \eta_J}.$$

This is to be compared with Vapnik's result, as proved in [37, page 138]:

Theorem 2.25 (Vapnik). *For any i.i.d. probability distribution $\mathbb{P}$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $\theta \in \Theta$, putting*

$$d_V = \log \bigl[\mathbb{P}(\mathfrak{N}_1) \bigr] + \log(4/\epsilon),$$

$$R(\theta) \leq r_1(\theta) + \frac{2 d_V}{N} + \sqrt{\frac{4 d_V r_1(\theta)}{N} + \frac{4 d_V^2}{N^2}}.$$

Recalling that we can choose $(\eta_j)_{j=1}^2$ such that $\eta_J = \frac{1}{10N}$ (which is negligeable by all means) and such that for any $N \leq 10^9$,

$$d_1''(\theta) \leq \mathbb{P} \bigl[\log(\mathfrak{N}_1) \,|\, (Z_i)_{i=1}^N \bigr] - \log(\epsilon) + 4.7,$$

we see that our complexity term is somehow more satisfactory than Vapnik's, since it is integrated outside the logarithm, with a little larger additional constant (remember that $\log(4) \simeq 1.4$, which is better than our 4.7, which could presumably be improved by working out a better sequence η_j, but not down to $\log(4)$). Our variance term is better, since we get $r_1(1-r_1)$ as we should, instead of only r_1. We also have $\dfrac{d_1''}{N}$ instead of $2\dfrac{d_V}{N}$, comming from the fact that we do not use any symmetrization trick.

Let us illustrate these bound on a numerical example (corresponding to a situation where the sample is noisy or the classification model is weak).

Let us assume that $N = 1000$, $\inf_{\Theta} r_1 = r_1(\widehat{\theta}) = 0.2$, that we are performing binary classification with a model with VC dimension not greater than $h = 10$, and that we work at level of confidence $\epsilon = 0.01$. Vapnik's theorem provides an upper bound for $R(\widehat{\theta})$ not smaller than 0.610, whereas Corollary 2.24 gives $R(\widehat{\theta}) \leq 0.461$ (using the bound $d_1'' \leq d_1' + 3.7$ when $N = 1000$). Now if we go for Theorem 2.20 and do not make a Gaussian approximation, we get $R(\widehat{\theta}) \leq 0.453$. It is interesting to remark that this bound is achieved for $\lambda = 1195 > N = 1000$. This explains why the Gaussian approximation in Vapnik's bound can be improved: for such a large value of λ, $\lambda r_1(\theta)$ does not behave like a Gaussian random variable.

Let us remind in conclusion that the best bound is provided by Theorem 2.17, giving $R(\widehat{\theta}) \leq 0.4211$, (that is approximately 2/3 of Vapnik's bound), for optimal values of $k = 15$, and of $\lambda = 1010$. This bound can be seen to take advantage of the fact that Bernoulli random variables are not Gaussian (its Gaussian approximation, Corollary 2.23, gives a bound $R(\theta) \simeq 0.4325$, still with an optimal $k = 15$), and of the fact that the optimal size of the shadow sample is significantly larger than the size of the observed sample. Moreover, Theorem 2.17 does not assume that the sample is i.i.d., but only that it is independent, thus generalizing Vapnik's bounds to inhomogeneous data (this will presumably be the case when data are collected from different places where the experimental conditions may not be expected to be the same, although they may reasonably be assumed to be independent). We would like also to emphasis that our little numerical example shows that Vapnik's bounds can be expected to be appropriate when dealing with moderate sample sizes. More sophisticated bounds obviously have a better asymptotic behaviour as shown in the first section. Nevertheless the numerical illustration of Theorem 1.18 given on page 30 suggests hat Vapnik's bounds are not doing so bad for small to medium ratios between the sample size and the dimension of the classification model (with local bounds, we could only get down to 0.332, although using a quite stronger dimension assumption).

We chose on purpose an example where it is non trivial to decide whether the chosen classifier does better than the 0.5 error rate of blind random classification. We think that this situation of weak learning is of practical interest, since "significant" weak learning rules may afterwards be aggregated or combined in various ways to achieve better classification rates.

3. Support Vector Machines

3.1. How to build them.

3.1.1. The canonical hyperplane. Support Vector Machines, of widely spread use and renown, were introduced by V. Vapkik [37]. Before introducing them, we will study as a prerequisite the separation of points by hyperplanes in a finite dimensional Euclidean space. Support Vector Machines perform the same kind of linear separation after an implicit change of pattern space. The preceding PAC-Bayesian results provide a fit framework to analyze their generalization properties.

We will deal in this section with the classification of points in $\mathbb{R}^d$ in two classes. Let $Z = (x_i, y_i)_{i=1}^N \in \bigl(\mathbb{R}^d \times \{-1, +1\}\bigr)^N$ be some set of labelled examples (called the training set hereafter). Let us split the set of indices $I = \{1, \dots, N\}$ according to the labels into two subsets

$$\begin{aligned} I_+ &= \{i \in I \, : y_i = +1\}, \\ I_- &= \{i \in I \, : y_i = -1\}. \end{aligned}$$

Let us then consider the set of admissible separating directions

$$A_Z = \bigl\{ w \in \mathbb{R}^d \, : \sup_{b \in \mathbb{R}} \inf_{i \in I} (\langle w, x_i \rangle - b) y_i \geq 1 \bigr\},$$

which can also be written as

$$A_Z = \bigl\{ w \in \mathbb{R}^d \, : \max_{i \in I_-} \langle w, x_i \rangle + 2 \leq \min_{i \in I_+} \langle w, x_i \rangle \bigr\}.$$

As it is easily seen, the optimal value of b for a fixed value of w, in other words the value of b which maximizes $\inf_{i \in I} (\langle w, x_i \rangle - b) y_i$, is equal to

$$b_w = \frac{1}{2} \Bigl[\max_{i \in I_-} \langle w, x_i \rangle + \min_{i \in I_+} \langle w, x_i \rangle \Bigr].$$

Lemma 3.1. *When $A_Z \neq \varnothing$, $\inf\{\|w\|^2 \, : w \in A_Z\}$ is reached for only one value w_Z of w.*

Proof. Let $w_0 \in A_Z$. The set $A_Z \cap \{w \in \mathbb{R}^d : \|w\| \leq \|w_0\|\}$ is a compact convex set and $w \mapsto \|w\|^2$ is strictly convex and therefore has a unique minimum on this set, which is also obviously its minimum on A_Z. $\square$

Definition 3.1. When $A_Z \neq \varnothing$, the training set Z is said to be linearly separable. The hyperplane

$$H = \{x \in \mathbb{R}^d \,:\, \langle w_Z, x\rangle - b_Z = 0\},$$

where

$$\begin{aligned} w_Z &= \arg\min\{\|w\| \,:\, w \in A_Z\}, \\ b_Z &= b_{w_Z}, \end{aligned}$$

is called the canonical separating hyperplane of the training set Z. The quantity $\|w_Z\|^{-1}$ is called the margin of the canonical hyperplane.

Note that as $\min_{i\in I_+}\langle w_Z, x_i\rangle - \max_{i \in I_-}\langle w_Z, x_i\rangle = 2$, the margin is also equal to half the distance between the projections on the direction w_Z of the positive and negative patterns.

3.1.2. Computation of the canonical hyperplane. Let us consider the convex hulls X_+ and X_- of the positive and negative patterns:

$$\begin{aligned} \mathcal{X}_+ &= \Big\{ \sum_{i\in I_+} \lambda_i x_i \,:\, \bigl(\lambda_i\bigr)_{i\in I_+} \in \mathbb{R}_+^{I_+}, \sum_{i\in I_+}\lambda_i = 1\Big\}, \\ \mathcal{X}_- &= \Big\{ \sum_{i\in I_-} \lambda_i x_i \,:\, \bigl(\lambda_i\bigr)_{i\in I_-} \in \mathbb{R}_+^{I_-}, \sum_{i\in I_-}\lambda_i = 1\Big\}. \end{aligned}$$

Let us introduce the closed convex set

$$\mathcal{V} = \mathcal{X}_+ - \mathcal{X}_- = \bigl\{ x_+ - x_- \,:\, x_+ \in \mathcal{X}_+, x_- \in \mathcal{X}_- \bigr\}.$$

As $v \mapsto \|v\|^2$ is strictly convex, with compact lower level sets, there is a unique vector v^* such that

$$\|v^*\|^2 = \inf_{v\in\mathcal{V}} \bigl\{ \|v\|^2 \,:\, v \in \mathcal{V}\bigr\}.$$

Lemma 3.2. *The set A_Z is non empty (i.e. the training set Z is linearly separable) if and only if $v^* \neq 0$. In this case*

$$w_Z = \frac{2}{\|v^*\|^2} v^*,$$

and the margin of the canonical hyperplane is equal to $\frac{1}{2}\|v^*\|$.

Proof. Let us assume first that $v^* = 0$, or equivalently that $\mathcal{X}_+ \cap \mathcal{X}_- \neq \varnothing$. As for any vector $w \in \mathbb{R}^d$,

$$\min_{i \in I_+} \langle w, x_i \rangle = \min_{x \in \mathcal{X}_+} \langle w, x \rangle,$$
$$\max_{i \in I_-} \langle w, x_i \rangle = \max_{x \in \mathcal{X}_-} \langle w, x \rangle,$$

we see that necessarily $\min_{i \in I_+} \langle w, x_i \rangle - \max_{i \in I_-} \langle w, x_i \rangle \leq 0$, which shows that w cannot be in A_Z and therefore that A_Z is empty.

Let us assume now that $v^* \neq 0$, or equivalently that $\mathcal{X}_+ \cap \mathcal{X}_- = \varnothing$. Let us put $w^* = \frac{2}{\|v^*\|^2} v^*$. Let us remark first that

$$\begin{aligned}
\min_{i \in I_+} \langle w^*, x_i \rangle - \max_{i \in I_-} \langle w^*, x_i \rangle &= \inf_{x \in \mathcal{X}_+} \langle w^*, x \rangle - \sup_{x \in \mathcal{X}_-} \langle w^*, x \rangle \\
&= \inf_{x_+ \in \mathcal{X}_+, x_- \in \mathcal{X}_-} \langle w^*, x_+ - x_- \rangle \\
&= \frac{2}{\|v^*\|^2} \inf_{v \in \mathcal{V}} \langle v^*, v \rangle.
\end{aligned}$$

Let us now prove that $\inf_{v \in \mathcal{V}} \langle v^*, v \rangle = \|v^*\|^2$. Some arbitrary $v \in \mathcal{V}$ being fixed, consider the function

$$\beta \mapsto \|\beta v + (1 - \beta) v^*\|^2 : [0, 1] \to \mathbb{R}.$$

By definition of v^*, it reaches its minimum value for $\beta = 0$, and therefore has a non negative derivative at this point. Computing this derivative, we find that $\langle v - v^*, v^* \rangle \geq 0$, as claimed. We have proved that

$$\min_{i \in I_+} \langle w^*, x_i \rangle - \max_{i \in I_-} \langle w^*, x_i \rangle = 2,$$

and therefore that $w^* \in A_Z$. On the other hand, any $w \in A_Z$ is such that

$$2 \leq \min_{i \in I_+} \langle w, x_i \rangle - \max_{i \in I_-} \langle w, x_i \rangle = \inf_{v \in \mathcal{V}} \langle w, v \rangle \leq \|w\| \inf_{v \in \mathcal{V}} \|v\| = \|w\| \, \|v^*\|.$$

This proves that $\|w^*\| = \inf \bigl\{ \|w\| \, : w \in A_Z \bigr\}$, and therefore that $w^* = w_Z$ as claimed. $\square$ One way to compute w_Z would be therefore to compute v^* by minimizing

$$\Bigl\{ \|\sum_{i \in I} \lambda_i y_i x_i\|^2 \, : (\lambda_i)_{i \in I} \in \mathbb{R}_+^I, \sum_{i \in I} \lambda_i = 2, \sum_{i \in I} y_i \lambda_i = 0 \Bigr\}.$$

Although this is a tractable quadratic programming problem, a direct computation of w_Z through the following proposition is usually prefered.

Proposition 3.3. *The canonical direction w_Z can be expressed as*

$$w_Z = \sum_{i=1}^{N} \alpha_i^* y_i x_i,$$

where $(\alpha_i^*)_{i=1}^N$ is obtained by minimizing

$$\inf\bigl\{F(\alpha) \,:\alpha \in \mathcal{A}\bigr\},$$

where

$$\mathcal{A} = \Bigl\{(\alpha_i)_{i\in I} \in \mathbb{R}_+^I, \sum_{i\in I} \alpha_i y_i = 0\Bigr\},$$

and

$$F(\alpha) = \Bigl\|\sum_{i\in I} \alpha_i y_i x_i\Bigr\|^2 - 2\sum_{i\in I} \alpha_i.$$

Proof. Let $w(\alpha) = \sum_{i\in I} \alpha_i y_i x_i$ and let $S(\alpha) = \frac{1}{2}\sum_{i\in I}\alpha_i$. We can express the function $F(\alpha)$ as $F(\alpha) = \|w(\alpha)\|^2 - 4S(\alpha)$. Moreover it is important to notice that for any $s \in \mathbb{R}_+$ $\{w(\alpha) \,:\alpha \in \mathcal{A}, S(\alpha) = s\} = s\mathcal{V}$. This shows that for any $s \in \mathbb{R}_+$, $\inf\{F(\alpha) \,:\alpha \in \mathcal{A}, S(\alpha) = s\}$ is reached and that for any $\alpha_s \in \{\alpha \in \mathcal{A} \,:\, S(\alpha) = s\}$ reaching this infimum, $w(\alpha_s) = sv^*$. As $s \mapsto s^2\|v^*\|^2 - 4s : \mathbb{R}_+ \to \mathbb{R}$ reaches its infimum for only one value s^* of s, namely at $s^* = \frac{2}{\|v^*\|^2}$, this shows that $F(\alpha)$ reaches its infimum on $\mathcal{A}$, and that for any $\alpha^* \in \mathcal{A}$ such that $F(\alpha^*) = \inf\{F(\alpha) \,:\, \alpha \in \mathcal{A}\}$, $w(\alpha^*) = \frac{2}{\|v^*\|^2}v^* = w_Z$. □

3.1.3. Support vectors.

Definition 3.2. The set of support vectors $\mathcal{S}$ is defined by

$$\mathcal{S} = \{x_i \,:\langle w_Z, x_i\rangle - b_Z = y_i\}.$$

Proposition 3.4. *Any α^* minimizing $F(\alpha)$ on $\mathcal{A}$ is such that*

$$\{x_i \,:\alpha_i^* > 0\} \subset \mathcal{S}.$$

This implies that the representation $w_Z = w(\alpha^*)$ involves in general only a limited number of non zero coefficients and that $w_Z = w_{Z'}$, where $Z' = \{(x_i, y_i) \,:x_i \in \mathcal{S}\}$.

Proof. Let us consider any given $i \in I_+$ and $j \in I_-$, such that $\alpha_i^* > 0$ and $\alpha_j^* > 0$ (there exists at least one such index in each set I_- and I_+, since the sum of the components of α^* on each of these sets are equal and since $\sum_{k \in I} \alpha_k^* > 0$). For any $t \in \mathbb{R}$, consider

$$\alpha_k(t) = \alpha_k^* + t\mathbb{1}(k \in \{i, j\}), \quad k \in I.$$

The vector $\alpha(t)$ is in $\mathcal{A}$ for any value of t in some neighborhood of 0, therefore $\frac{\partial}{\partial t}_{|t=0} F\bigl[\alpha(t)\bigr] = 0$. Computing this derivative, we find that

$$y_i \langle w(\alpha^*), x_i \rangle + y_j \langle w(\alpha^*), x_j \rangle = 2.$$

As $y_i = -y_j$, this can also be written as

$$y_i \bigl[\langle w(\alpha^*), x_i \rangle - b_Z\bigr] + y_j \bigl[\langle w(\alpha^*), x_j \rangle - b_Z\bigr] = 2.$$

As $w(\alpha^*) \in A_Z$,

$$y_k \bigl[\langle w(\alpha^*), x_k \rangle - b_Z\bigr] \geq 1, \qquad k \in I,$$

which implies necessarily as claimed that

$$y_i \bigl[\langle w(\alpha^*), x_i \rangle - b_Z\bigr] = y_j \bigl[\langle w(\alpha^*), x_j \rangle - b_Z\bigr] = 1.$$

□

3.1.4. The non separable case. In the case when the training set $Z = (x_i, y_i)_{i=1}^N$ is not linearly separable, we can define a noisy canonical hyperplane as follows. We can choose $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ to minimize

$$C(w, b) = \sum_{i=1}^N \bigl[1 - \bigl(\langle w, x_i \rangle - b\bigr) y_i\bigr]_+ + \tfrac{1}{2}\|w\|^2, \tag{3.1}$$

where for any real number r, $r_+ = \max\{r, 0\}$ is the positive part of r.

Theorem 3.5. *Let us introduce the dual criterion*

$$F(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2}\Bigl\| \sum_{i=1}^N y_i \alpha_i x_i \Bigr\|^2$$

and the domain $\mathcal{A}' = \Bigl\{ \alpha \in \mathbb{R}_+^N : \alpha_i \leq 1, i = 1, \dots, N, \sum_{i=1}^N y_i \alpha_i = 0 \Bigr\}$. *Let* $\alpha^* \in \mathcal{A}'$ *be such that* $F(\alpha^*) = \sup_{\alpha \in \mathcal{A}'} F(\alpha)$. *Let* $w^* = \sum_{i=1}^N y_i \alpha_i^* x_i$. *There is a threshold* b^* *(whose construction will be detailed in the proof), such that*

$$C(w^*, b^*) = \inf_{w \in \mathbb{R}^d, b \in \mathbb{R}} C(w, b).$$

Corollary 3.6. (scaled criterion) *For any positive real parameter* λ *let us consider the criterion*

$$C_\lambda(w,b) = \lambda^2 \sum_{i=1}^N \bigl[1 - (\langle w, x_i \rangle - b) y_i\bigr]_+ + \tfrac{1}{2}\|w\|^2$$

and the domain $\mathcal{A}'_\lambda = \left\{ \alpha \in \mathbb{R}_+^N : \alpha_i \leq \lambda^2, i = 1, \dots, N, \sum_{i=1}^N y_i \alpha_i = 0 \right\}$. *For any solution* α^* *of the minimization problem* $F(\alpha^*) = \sup_{\alpha \in \mathcal{A}'_\lambda} F(\alpha)$, *the vector* $w^* = \sum_{i=1}^N y_i \alpha_i^* x_i$ *is such that*

$$\inf_{b \in \mathbb{R}} C_\lambda(w^*, b) = \inf_{w \in \mathbb{R}^d, b \in \mathbb{R}} C_\lambda(w,b).$$

Let us remark that in the separable case, the scaled criterion is minimized by the canonical hyperplane for λ large enough. This extension of the canonical hyperplane computation in dual space is often called *the box constraint*, for obvious reasons.

Proof. The corollary is a straightforward consequence of the scale property $C_\lambda(w,b,x) = \lambda^2 C(\lambda^{-1} w, b, \lambda x)$, where we have made the dependence of the criterion in $x \in \mathbb{R}^{dN}$ explicit. Let us come now to the proof of the theorem.

The minimization of $C(w,b)$ can be performed in dual space extending the couple of parameters (w,b) to $\overline{w} = (w,b,\gamma) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+^N$ and introducing the dual multipliers $\alpha \in \mathbb{R}_+^N$ and the criterion

$$G(\alpha, \overline{w}) = \sum_{i=1}^N \gamma_i + \sum_{i=1}^N \alpha_i \bigl\{ \bigl[1 - (\langle w, x_i \rangle - b) y_i\bigr] - \gamma_i \bigr\} + \tfrac{1}{2}\|w\|^2.$$

We see that

$$C(w,b) = \inf_{\gamma \in \mathbb{R}_+^N} \sup_{\alpha \in \mathbb{R}_+^N} G\bigl[\alpha, (w,b,\gamma)\bigr],$$

and therefore, putting $\overline{\mathcal{W}} = \bigl\{ (w,b,\gamma) : w \in \mathbb{R}^d, b \in \mathbb{R}, \gamma \in \mathbb{R}_+^N \bigr\}$, we are led to solve the minimization problem

$$G(\alpha_*, \overline{w}_*) = \inf_{\overline{w} \in \overline{\mathcal{W}}} \sup_{\alpha \in \mathbb{R}_+^N} G(\alpha, \overline{w}),$$

whose solution $\overline{w}_* = (w_*, b_*, \gamma_*)$ is such that $C(\overline{w}_*, b_*) = \inf_{(w,b) \in \mathbb{R}^{d+1}} C(w,b)$, according to the preceding identity. As for any value of $\alpha' \in \mathbb{R}_+^N$,

$$\inf_{\overline{w} \in \overline{\mathcal{W}}} \sup_{\alpha \in \mathbb{R}_+^N} G(\alpha, \overline{w}) \geq \inf_{\overline{w} \in \overline{\mathcal{W}}} G(\alpha', \overline{w}),$$

it is immediately seen that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha,\overline{w}) \geq \sup_{\alpha\in\mathbb{R}_+^N} \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha,\overline{w}).$$

We are going to show that there is no duality gap, meaning that this inequality is indeed an equality. More importantly, we will do so by exhibiting a saddle point, which, solving the dual minimization problem will also solve the original one.

Let us first make explicit the solution of the dual problem (the interest of this dual problem precisely lies in the fact that it can more easily be solved explicitly). Introducing the admissible set of values of α,

$$\mathcal{A}' = \Bigl\{\alpha \in \mathbb{R}^N : 0 \leq \alpha_i \leq 1, i = 1,\dots,N, \sum_{i=1}^N y_i\alpha_i = 0\Bigr\},$$

it is elementary to check that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha,\overline{w}) = \begin{cases} \inf_{w\in\mathbb{R}^d} G\bigl[\alpha,(w,0,0)\bigr], & \alpha\in\mathcal{A}', \\ -\infty, & \text{otherwise.}\end{cases}$$

As

$$G\bigl[\alpha,(w,0,0)\bigr] = \tfrac{1}{2}\|w\|^2 + \sum_{i=1}^N \alpha_i\bigl(1 - \langle w, x_i\rangle y_i\bigr),$$

we see that $\inf_{w\in\mathbb{R}^d} G\bigl[\alpha,(w,0,0)\bigr]$ is reached at

$$w_\alpha = \sum_{i=1}^N y_i\alpha_i x_i.$$

This proves that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha,\overline{w}) = F(\alpha).$$

The continuous map $\alpha \mapsto \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha,\overline{w})$ reaches a (non necessarily unique) maximum α^* on the compact convex set $\mathcal{A}'$. We are now going to exhibit a choice of $\overline{w}^* \in \overline{\mathcal{W}}$ such that $(\alpha^*,\overline{w}^*)$ is a *saddle point*. This means that we are going to show that

$$G(\alpha^*,\overline{w}^*) = \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha^*,\overline{w}) = \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha,\overline{w}^*).$$

It will imply that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^d} G(\alpha,\overline{w}) \leq \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha,\overline{w}^*) = G(\alpha^*,\overline{w}^*)$$

on the one hand and that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^d} G(\alpha,\overline{w}) \geq \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha^*,\overline{w}) = G(\alpha^*,\overline{w}^*)$$

on the other hand, proving that

$$G(\alpha^*,\overline{w}^*) = \inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha,\overline{w})$$

as required.
Construction of $\overline{w}^*$.

- Let us put $w^* = w_{\alpha^*}$.
- If there is $j \in \{1,\dots,N\}$ such that $0 < \alpha_j^* < 1$, let us put

  $$b^* = \langle x_j, w^* \rangle - y_j.$$

  Otherwise, let us put

  $$b^* = \sup\{\langle x_i, w^*\rangle - 1 : \alpha_i^* > 0, y_i = +1, i = 1,\dots,N\}.$$

- Let us then put

  $$\gamma_i^* = \begin{cases} 0, & \alpha_i^* < 1,\\ 1 - (\langle w^*, x_i\rangle - b^*)y_i, & \alpha_i^* = 1. \end{cases}$$

If we can prove that

$$1 - (\langle w^*, x_i\rangle - b^*)y_i \begin{cases} \leq 0, & \alpha_i^* = 0,\\ = 0, & 0 < \alpha_i^* < 1,\\ \geq 0, & \alpha_i^* = 1, \end{cases} \tag{3.2}$$

it will show that $\gamma^* \in \mathbb{R}_+^N$ and therefore that $\overline{w}^* = (w^*, b^*, \gamma^*) \in \overline{\mathcal{W}}$. It will also show that

$$G(\alpha,\overline{w}^*) = \sum_{i=1}^N \gamma_i^* + \sum_{i,\alpha_i^*=0} \alpha_i \bigl[1 - (\langle \overline{w}^*, x_i\rangle - b^*)y_i\bigr] + \tfrac{1}{2}\|\overline{w}^*\|^2,$$

proving that $G(\alpha^*, \overline{w}^*) = \sup_{\alpha \in \mathbb{R}_+^N} G(\alpha, \overline{w}^*)$. As obviously $G(\alpha^*, \overline{w}^*) = G\bigl[\alpha^*, (w^*, 0, 0)\bigr]$, we already know that $G(\alpha^*, \overline{w}^*) = \inf_{\overline{w} \in \overline{\mathcal{W}}} G(\alpha^*, \overline{w})$. This will show that $(\alpha^*, \overline{w}^*)$ is the saddle point we were looking for, thus ending the proof of the theorem.

Proof of equation (3.2): Let us deal first with the case when there is $j \in \{1, \dots, N\}$ such that $0 < \alpha_j^* < 1$.

For any $i \in \{1, \dots, N\}$ such that $0 < \alpha_i^* < 1$, there is $\epsilon > 0$ such that for any $t \in (-\epsilon, \epsilon)$, $\alpha^* + t y_i e_i - t y_j e_j \in \mathcal{A}'$, where $(e_k)_{k=1}^N$ is the canonical base of $\mathbb{R}^N$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + t y_i e_i - t y_j e_j) = 0$. Computing this derivative, we obtain

$$\begin{aligned} \frac{\partial}{\partial t}_{|t=0} F(\alpha^* + t y_i e_i - t y_j e_j) &= y_i - \langle w^*, x_i \rangle + \langle w^*, x_j \rangle - y_j \\ &= y_i \bigl[1 - \bigl(\langle w, x_i \rangle - b^*\bigr) y_i\bigr]. \end{aligned}$$

Thus $1 - \bigl(\langle w, x_i \rangle - b^*\bigr) y_i = 0$, as required. This shows also that the definition of b^* does not depend on the choice of j such that $0 < \alpha_j^* < 1$.

For any $i \in \{1, \dots, N\}$ such that $\alpha_i^* = 0$, there is $\epsilon > 0$ such that for any $t \in (0, \epsilon)$, $\alpha^* + t e_i - t y_i y_j e_j \in \mathcal{A}'$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + t e_i - t y_i y_j e_j) \leq 0$, showing that $1 - \bigl(\langle w^*, x_i \rangle - b^*\bigr) y_i \leq 0$ as required.

For any $i \in \{1, \dots, N\}$ such that $\alpha_i^* = 1$, there is $\epsilon > 0$ such that $\alpha^* - t e_i + t y_i y_j e_j \in \mathcal{A}'$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* - t e_i + t y_i y_j e_j) \leq 0$, showing that $1 - \bigl(\langle w^*, x_i \rangle - b^*\bigr) y_i \geq 0$ as required. This ends to prove that $(\alpha^*, \overline{w}^*)$ is a saddle point in this case.

Let us deal now with the case where $\alpha^* \in \{0, 1\}^N$. If we are not in the trivial case where the vector $(y_i)_{i=1}^N$ is constant, the case $\alpha^* = 0$ is ruled out. Indeed, in this case, considering $\alpha^* + t e_i + t e_j$, where $y_i y_j = -1$, we would get the contradiction $2 = \frac{\partial}{\partial t}_{|t=0} F(\alpha^* + t e_i + t e_j) \leq 0$.

Thus there are values of j such that $\alpha_j^* = 1$, and since $\sum_{i=1}^N \alpha_i y_i = 0$, both classes are present in the set $\{j : \alpha_j^* = 1\}$.

Now for any $i, j \in \{1, \dots, N\}$ such that $\alpha_i^* = \alpha_j^* = 1$ and such that $y_i = +1$ and $y_j = -1$, $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* - t e_i - t e_j) = -2 + \langle w^*, x_i \rangle - \langle w^*, x_j \rangle \leq 0$. Thus

$$\sup\{\langle w^*, x_i \rangle - 1 : \alpha_i^* = 1, y_i = +1\} \leq \inf\{\langle w^*, x_j \rangle + 1 : \alpha_j^* = 1, y_j = -1\},$$

showing that

$$1 - \bigl(\langle w^*, x_k \rangle - b^*\bigr) y_k \geq 0, \alpha_k^* = 1.$$

Eventually, for any i such that $\alpha_i^* = 0$, for any j such that $\alpha_j^* = 1$ and $y_j = y_i$

$$\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + te_i - te_j) = y_i \langle w^*, x_i - x_j \rangle \leq 0,$$

showing that $1 - \bigl(\langle w^*, x_i \rangle - b^* \bigr) y_i \leq 0$. This ends to prove that $(\alpha^*, \overline{w}^*)$ is in all circumstances a saddle point.

3.1.5. Support Vector Machines.

DEFINITION 3.3. The symmetric measurable kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is said to be positive (or more precisely positive semi-definite) if for any $n \in \mathbb{N}$, any $(x_i)_{i=1}^n \in \mathcal{X}^n$,

$$\inf_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \sum_{j=1}^n \alpha_i K(x_i, x_j) \alpha_j \geq 0.$$

Let $Z = (x_i, y_i)_{i=1}^N$ be some training set. Let us consider as previously

$$\mathcal{A} = \Bigl\{ \alpha \in \mathbb{R}_+^N : \sum_{i=1}^N \alpha_i y_i = 0 \Bigr\}.$$

Let

$$F(\alpha) = \sum_{i=1}^N \sum_{j=1}^N \alpha_i y_i K(x_i, x_j) y_j \alpha_j - 2 \sum_{i=1}^N \alpha_i.$$

DEFINITION 3.4. Let K be a positive symmmetric kernel. The training set Z is said to be K-separable if

$$\inf \bigl\{ F(\alpha) : \alpha \in \mathcal{A} \bigr\} > -\infty.$$

LEMMA 3.7. *When Z is K-separable,* $\inf \{ F(\alpha) : \alpha \in \mathcal{A} \}$ *is reached.*

PROOF. Consider the training set $Z' = (x_i', y_i)_{i=1}^N$, where

$$x_i' = \left\{ \left[\Bigl\{ K(x_k, x_\ell) \Bigr\}_{k=1, \ell=1}^{N \quad N} \right]^{1/2} (i, j) \right\}_{j=1}^N \in \mathbb{R}^N.$$

We see that $F(\alpha) = \Bigl\| \sum_{i=1}^N \alpha_i y_i x_i' \Bigr\|^2 - 2 \sum_{i=1}^N \alpha_i$. We have proved in the previous section that Z' is linearly separable if and only if $\inf \{ F(\alpha) : \alpha \in \mathcal{A} \} > -\infty$, and that the infimum is reached in this case. $\square$

Proposition 3.8. *Let K be a symmetric positive kernel and let $Z = (x_i, y_i)_{i=1}^N$ be some K-separable training set. Let $\alpha^* \in \mathcal{A}$ be such that $F(\alpha^*) = \inf\{F(\alpha) : \alpha \in \mathcal{A}\}$. Let*

$$I_-^* = \{i \in \mathbb{N} : 1 \leq i \leq N, y_i = -1, \alpha_i^* > 0\}$$
$$I_+^* = \{i \in \mathbb{N} : 1 \leq i \leq N, y_i = +1, \alpha_i^* > 0\}$$
$$b^* = \frac{1}{2}\Big\{\sum_{j=1}^N \alpha_j^* y_j K(x_j, x_{i_-}) + \sum_{j=1}^N \alpha_j^* y_j K(x_j, x_{i_+})\Big\}, \qquad i_- \in I_-^*, i_+ \in I_+^*,$$

where the value of b^* does not depend on the choice of i_- and i_+. The classification rule $f : \mathcal{X} \to \mathcal{Y}$ defined by the formula

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^N \alpha_i^* y_i K(x_i, x) - b^*\right)$$

is independent of the choice of α^* and is called the support vector machine defined by K and Z. The set $\mathcal{S} = \{x_j : \sum_{i=1}^N \alpha_i^* y_i K(x_i, x_j) - b^* = y_j\}$ is called the set of support vectors. For any choice of α^*, $\{x_i : \alpha_i^* > 0\} \subset \mathcal{S}$.

An important consequence of this proposition is that the support vector machine defined by K and Z is also the support vector machine defined by K and $Z' = \{(x_i, y_i) : \alpha_i^* > 0, 1 \leq i \leq N\}$, since this restriction of the index set contains the value α^* where the minimum of F is reached.

Proof. The independence from the choice of α^*, which is not necessarily unique, is seen as follows. Let $(x_i)_{i=1}^N$ and $x \in \mathcal{X}$ be fixed. Let us put for ease of notations $x_{N+1} = x$. Let M be the $(N+1) \times (N+1)$ symmetric semi-definite matrix defined by $M(i,j) = K(x_i, x_j)$, $i = 1, \dots, N+1$, $j = 1, \dots, N+1$. Let us consider the mapping $\Psi : \{x_i : i = 1, \dots, N+1\} \to \mathbb{R}^{N+1}$ defined by

$$\Psi(x_i) = \big[M^{1/2}(i,j)\big]_{j=1}^{N+1} \in \mathbb{R}^{N+1}. \tag{3.3}$$

Let us consider the training set $Z' = \big[\Psi(x_i), y_i\big]_{i=1}^N$. Then Z' is linearly separable,

$$F(\alpha) = \Big\|\sum_{i=1}^N \alpha_i y_i \Psi(x_i)\Big\|^2 - 2\sum_{i=1}^N \alpha_i,$$

and we have proved that for any choice of $\alpha^* \in \mathcal{A}$ minimizing $F(\alpha)$, $w_{Z'} = \sum_{i=1}^N \alpha_i^* y_i \Psi(x_i)$. Thus the support vector machine defined by K and Z can also be expressed by the formula

$$f(x) = \operatorname{sign}\Big[\langle w_{Z'}, \Psi(x)\rangle - b_{Z'}\Big]$$

which does not depend on α^*. The definition of $\mathcal{S}$ is such that $\Psi(\mathcal{S})$ is the set of support vectors defined in the linear case, where its stated property has already been prooved. $\square$

We can in the same way use the box constraint and show that any solution $\alpha^* \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i \leq \lambda^2, i = 1, \dots, N\}$ minimizes

$$\inf_{b \in \mathbb{R}} \lambda^2 \sum_{i=1}^N \Biggl[1 - \Biggl(\sum_{j=1}^N y_j \alpha_j K(x_j, x_i) - b \Biggr) y_i \Biggr]_+ \\ + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j). \quad (3.4)$$

3.1.6. Building kernels. The results of this section (except the last one) are drawned from [18]. We have no reference for the last proposition of this section, although we believe it is well known. We include them for the convenience of the reader.

Proposition 3.9. *Let K_1 and K_2 be positive symmetric kernels on $\mathcal{X}$. Then for any $a \in \mathbb{R}_+$*

$$\begin{aligned} (aK_1 + K_2)(x, x') &\overset{\text{def}}{=} aK_1(x, x') + K_2(x, x') \\ \textit{and } (K_1 \cdot K_2)(x, x') &\overset{\text{def}}{=} K_1(x, x') K_2(x, x') \end{aligned}$$

are also positive symmetric kernels. Moreover, for any measurable function $g : \mathcal{X} \to \mathbb{R}$, $K_g(x, x') \overset{\text{def}}{=} g(x) g(x')$ is also a positive symmetric kernel.

Proof. It is enough to prove the proposition in the case when $\mathcal{X}$ is finite and kernels are just ordinary symmetric matrices. Thus we can assume without loss of generality that $\mathcal{X} = \{1, \dots, n\}$. Then for any $\alpha \in \mathbb{R}^N$, using usual matrix notations,

$$\begin{aligned} \langle \alpha, (aK_1 + K_2)\alpha \rangle &= a \langle \alpha, K_1 \alpha \rangle + \langle \alpha, K_2 \alpha \rangle \geq 0, \\ \langle \alpha, (K_1 \cdot K_2) \alpha \rangle &= \sum_{i,j} \alpha_i K_1(i,j) K_2(i,j) \alpha_j \\ &= \sum_{i,j,k} \alpha_i K_1^{1/2}(i,k) K_1^{1/2}(k,j) K_2(i,j) \alpha_j \\ &= \sum_k \underbrace{\sum_{i,j} \bigl[K_1^{1/2}(k,i) \alpha_i \bigr] K_2(i,j) \bigl[K_1^{1/2}(k,j) \alpha_j \bigr]}_{\geq 0} \geq 0, \end{aligned}$$

$$\langle \alpha, K_g \alpha \rangle = \sum_{i,j} \alpha_i g(i) g(j) \alpha_j = \left(\sum_i \alpha_i g(i) \right)^2 \geq 0.$$

□

Proposition 3.10. *Let K be some positive symmetric kernel on $\mathcal{X}$. Let $p : \mathbb{R} \rightarrow \mathbb{R}$ be a polynomial with positive coefficients. Let $g : \mathcal{X} \rightarrow \mathbb{R}^d$ be a measurable function. Then*

$$\begin{aligned} p(K)(x,x') &\overset{def}{=} p\bigl[K(x,x')\bigr], \\ \exp(K)(x,x') &\overset{def}{=} \exp\bigl[K(x,x')\bigr] \\ \text{and } G_g(x,x') &\overset{def}{=} \exp\bigl(-\|g(x) - g(x')\|^2\bigr) \end{aligned}$$

are all positive symmetric kernels.

Proof. The first assertion is a direct consequence of the previous proposition. The second one comes from the fact that the exponential function is the pointwise limit of a sequence of polynomial functions with positive coefficients. The third one is seen from the second one and the decomposition

$$G_g(x,x') = \Bigl[\exp\bigl(-\|g(x)\|^2\bigr) \exp\bigl(-\|g(x')\|^2\bigr)\Bigr] \exp\bigl[2\langle g(x), g(x')\rangle\bigr]$$

□

Proposition 3.11. *With the notations of the previous proposition,* any *training set $Z = (x_i, y_i)_{i=1}^N \in \bigl(\mathcal{X} \times \{-1,+1\}\bigr)^N$ is G_g-separable as soon as $g(x_i)$, $i = 1, \dots, N$ are distinct points of $\mathbb{R}^d$.*

Proof. It is clearly enough to prove the case when $\mathcal{X} = \mathbb{R}^d$ and g is the identity. Let us consider some other generic point $x_{N+1} \in \mathbb{R}^d$ and define Ψ as in (3.3). It is enough to prove that $\Psi(x_1), \dots, \Psi(x_N)$ are affine independent, since the simplex, and therefore any affine independent set of points can be shattered by affine half-spaces. Let us assume that $(x_1, \dots, x_N)$ are affine dependent, this means that for some $(\lambda_1, \dots, \lambda_N) \neq 0$ such that $\sum_{i=1}^N \lambda_i = 0$,

$$\sum_{i=1}^N \sum_{j=1}^N \lambda_i G(x_i, x_j) \lambda_j = 0.$$

Thus, $(\lambda_i)_{i=1}^{N+1}$, where we have put $\lambda_{N+1} = 0$ is in the kernel of the symmetric positive semi-definite matrix $G(x_i, x_j)_{i,j \in \{1,\dots,N+1\}}$. Therefore

$$\sum_{i=1}^{N} \lambda_i G(x_i, x_{N+1}) = 0,$$

for any $x_{N+1} \in \mathbb{R}^d$. This would mean that the functions $x \mapsto \exp(-\|x - x_i\|^2)$ are linearly dependent, which can be easily proved to be false. Indeed, let $n \in \mathbb{R}^d$ be such that $\|n\| = 1$ and $\langle n, x_i \rangle$, $i = 1, \dots, N$ are distinct (such a vector exists, because it has to be outside the union of a finite number of hyperplanes, which is of zero Lebesgue measure on the sphere). Let us assume for a while that for some $(\lambda_i)_{i=1}^N \in \mathbb{R}^N$, for any $x \in \mathbb{R}^d$,

$$\sum_{i=1}^{N} \lambda_i \exp(-\|x - x_i\|^2) = 0.$$

Considering $x = tn$, for $t \in \mathbb{R}$, we would get

$$\sum_{i=1}^{N} \lambda_i \exp(2t\langle n, x_i \rangle - \|x_i\|^2) = 0, \qquad t \in \mathbb{R}.$$

Letting t go to infinity, we see that this is only possible if $\lambda_i = 0$ for all values of i. □

3.2. Bounds for Support Vector Machines.

3.2.1. Compression scheme bounds. We can use Support Vector Machines in the framework of compression schemes and apply Theorem 2.17 on page 114. More precisely, given some positive symmetric kernel K on $\mathcal{X}$, we may consider for any training set $Z' = (x'_i, y'_i)_{i=1}^h$ the classifier $\hat{f}_{Z'} : \mathcal{X} \to \mathcal{Y}$ which is equal to the Support Vector Machine defined by K and Z' whenever Z' is K-separable, and which is equal to some constant classification rule otherwise (we take this convention to stick to the framework described on page 105, we will only use $\hat{f}_{Z'}$ in the K-separable case, so this extension of the definition is just a matter of presentation). In the application of Theorem 2.17 in the case when the observed sample $(X_i, Y_i)_{i=1}^N$ is K-separable, a natural (if not always optimal) choice of Z' is to choose for (x'_i) the set of support vectors defined by $Z = (X_i, Y_i)_{i=1}^N$ and to choose for (y'_i) the corresponding values of Y. This is justified by the fact that $\hat{f}_Z = \hat{f}_{Z'}$, as shown in Proposition 3.8 (page 131). In the case when Z is not K-separable,

we can train a Support Vector Machine with the box constraint, then remove all the errors to obtain a K-separable subsample $Z' = \{(X_i, Y_i) : \alpha_i^* < \lambda^2, 1 \leq i \leq N\}$, (using the same notations as in equation (3.4) on page 132) and then consider its support vectors as the compression set. Still using the notations of page 132, this means we have to compute successively $\alpha^* \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i \leq \lambda^2\}$, and $\alpha^{**} \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i = 0 \text{ when } \alpha_i^* = \lambda^2\}$, to keep eventually the compression set indexed by $J = \{i : 1 \leq i \leq N, \alpha_i^{**} > 0\}$, and the corresponding Support Vector Machine $\widehat{f}_J$. Different values of λ can be used at this stage, producing different candidate compression sets : when λ increases, the number of errors should decrease, on the other hand when λ decreases, the margin $\|w\|^{-1}$ of the separable subset Z' increases, supporting the hope for a smaller set of support vectors, thus we can use λ to monitor the number of errors on the training set we accept from the compression scheme. As we can use whatever heuristic we want while selecting the compression set, we can also try to threshold in the previous construction α_i^{**} at different levels $\eta \geq 0$, to produce candidate compression sets $J_\eta = \{i : 1 \leq i \leq N, \alpha_i^{**} > \eta\}$ of various sizes.

As the size $|J|$ of the compression set is random in this construction, we have to use a version of Theorem 2.17 (page 114) which handles compression sets of arbitrary sizes. This is done by choosing for each k a k-partially exchangeable posterior distribution π_k which weights the compression sets of all dimensions. We immediately see that we can choose π_k such that $-\log\bigl[\pi_k(\Delta_k(J))\bigr] \leq \log\bigl[|J|(|J|+1)\bigr] + |J|\log\Bigl[\frac{(k+1)eN}{|J|}\Bigr]$.

If we observe the shadow sample patterns, and if computer resources permit, we can of course use more elaborate bounds than Theorem 2.17, such as the transductive correspondent to Theorem 1.24 (page 39) (where we may consider the submodels made of all the compression sets of the same size). Theorems based on relative bounds, such as Theorem 1.59 (page 88) can also be used. Gibbs distributions can be approximated by Monte Carlo techniques, where a Markov chain with the proper invariant measure consists in suitable local perturbations of the compression set.

Let us mention also that the use of compression schemes based on Support Vector Machines can be tailored to perform some kind of *feature aggregation*. Imagine that the kernel K is defined as the scalar product in $L_2(\pi)$, where $\pi \in \mathcal{M}_+^1(\Theta)$. More precisely let us consider for some set of soft classification rules $\bigl\{f_\theta : \mathcal{X} \rightarrow \mathbb{R}\,; \theta \in \Theta\bigr\}$ the kernel

$$K(x, x') = \int_{\theta \in \Theta} f_\theta(x) f_\theta(x') \pi(d\theta).$$

In this setting, the Support Vector Machine applied to the training set $Z =$

$(x_i, y_i)_{i=1}^N$ has the form

$$f_Z(x) = \operatorname{sign}\left(\int_{\theta \in \Theta} f_\theta(x) \sum_{i=1}^N y_i \alpha_i f_\theta(x_i) \pi(d\theta) - b \right)$$

and, may it be too burdening to compute, we can replace it with some finite approximation

$$\widetilde{f}_Z(x) = \operatorname{sign}\left(\sum_{k=1}^m f_{\theta_k}(x) w_k - b \right),$$

where the set $\{\theta_k,\, k = 1, \dots, m\}$ and the weights $\{w_k,\, k = 1, \dots, m\}$ are computed in some suitable way from $Z' = (x_i, y_i)_{i, \alpha_i > 0}$, the set of support vectors of f_Z. For instance, we can draw $\{\theta_k,\, k = 1, \dots, m\}$ at random according to the probability distribution proportional to

$$\left| \sum_{i=1}^N y_i \alpha_i f_\theta(x_i) \right| \pi(d\theta),$$

define the weights w_k by

$$w_k = \operatorname{sign}\left(\sum_{i=1}^N y_i \alpha_i f_{\theta_k}(x_i) \right) \int_{\theta \in \Theta} \left| \sum_{i=1}^N y_i \alpha_i f_\theta(x_i) \right| \pi(d\theta),$$

and choose the smallest value of m for which this approximation still classifies Z' without errors. Let us remark that we have built $\widetilde{f}_Z$ in such a way that

$$\lim_{m \to +\infty} \widetilde{f}_Z(x_i) = f_Z(x_i) = y_i, \quad \text{a.s.}$$

for any support index i such that $\alpha_i > 0$.

Alternatively, given Z', we can select a finite set of features $\Theta' \subset \Theta$ such that Z' is $K_{\Theta'}$ separable, where $K_{\Theta'}(x, x') = \sum_{\theta \in \Theta'} f_\theta(x) f_\theta(x')$ and consider the Support Vector Machines $f_{Z'}$ built with the kernel $K_{\Theta'}$. As soon as Θ' is chosen as a function of Z' only, Theorem 2.17 (page 114) applies and provides some level of confidence for the risk of $f_{Z'}$.

3.2.2. The Vapnik Cervonenkis dimension of a family of subsets. Let us consider some set X and some set $S \subset \{0,1\}^X$ of subsets of X. Let $h(S)$ be the VC dimension of S, defined as

$$h(S) = \max\{|A| : A \text{ finite and } A \cap S = \{0,1\}^A\},$$

where by definition $A \cap S = \{A \cap B : B \in S\}$. Let us notice that this definition does not depend on the choice of the reference set X. Indeed X can be chosen to be $\bigcup S$, the union of all the sets in S or any bigger set. Let us notice also that for any set B, $h(B \cap S) \leq h(S)$, the reason being that $A \cap (B \cap S) = B \cap (A \cap S)$.

This notion of VC dimension is useful because it can, as we will see about Support Vector Machines, be computed in some important special cases. Let us prove here as an illustration that $h(S) = d + 1$ when $X = \mathbb{R}^d$ and S is made of all the half spaces :

$$S = \{A_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}, \text{ where } A_{w,b} = \{x \in X : \langle w, x\rangle \geq b\}.$$

Proposition 3.12. *With the previous notations, $h(S) = d + 1$.*

Proof. Let $(e_i)_{i=1}^{d+1}$ be the canonical base of $\mathbb{R}^{d+1}$, and let X be the affine subspace it generates, which can be identified with $\mathbb{R}^d$. For any $(\epsilon_i)_{i=1}^{d+1} \in \{-1, +1\}^{d+1}$, let $w = \sum_{i=1}^{d+1} \epsilon_i e_i$ and $b = 0$. The half space $A_{w,b} \cap X$ is such that $\{e_i \,; i = 1, \dots, d + 1\} \cap (A_{w,b} \cap X) = \{e_i \,; \epsilon_i = +1\}$. This proves that $h(S) \geq d + 1$.

To prove that $h(S) \leq d+1$, we have to show that for any set $A \subset \mathbb{R}^d$ of size $|A| = d+2$, there is $B \subset A$ such that $B \not\in (A\cap S)$. This will obviously be the case if the convex hulls of B and $A \backslash B$ have a non empty intersection : indeed if a hyperplane separates two sets of points, it also separates their convex hulls. As $|A| > d + 1$, A is affine dependent : there is $(\lambda_x)_{x\in A} \in \mathbb{R}^{d+2} \setminus \{0\}$ such that $\sum_{x\in A} \lambda_x x = 0$ and $\sum_{x\in A} \lambda_x = 0$. The set $B = \{x \in A : \lambda_x > 0\}$ is non-empty, as well as its complement $A \setminus B$, because $\sum_{x\in A} \lambda_x = 0$ and $\lambda \neq 0$. Moreover $\sum_{x\in B} \lambda_x = \sum_{x\in A\setminus B} -\lambda_x > 0$. The relation

$$\frac{1}{\sum_{x\in B} \lambda_x} \sum_{x\in B} \lambda_x x = \frac{1}{\sum_{x\in B} \lambda_x} \sum_{x\in A\setminus B} -\lambda_x x$$

shows that the convex hulls of B and $A \setminus B$ have a non void intersection. □

Let us introduce the function of two integers

$$\Phi_n^h = \sum_{k=0}^{h} \binom{n}{k}$$

Let us notice that Φ can alternatively be defined by the relations :

$$\Phi_n^h = \begin{cases} 2^n & \text{when } n \leq h, \\ \Phi_{n-1}^{h-1} + \Phi_{n-1}^h & \text{when } n > h. \end{cases}$$

THEOREM 3.13. *Whenever* $\bigcup S$ *is finite,*

$$|S| \leq \Phi\left(\left|\bigcup S\right|, h(S)\right).$$

THEOREM 3.14. *For any* $h \leq n$,

$$\Phi_n^h \leq \exp\bigl(nH(\tfrac{h}{n})\bigr) \leq \exp\bigl[h\bigl(\log(\tfrac{n}{h}) + 1\bigr)\bigr],$$

where $H(p) = -p\log(p) - (1-p)\log(1-p)$ *is the Shannon entropy of the Bernoulli distribution with parameter* p.

PROOF OF THEOREM 3.13. Let us prove this theorem by induction on $|\bigcup S|$. It is easy to check that it holds true when $|\bigcup S| = 1$. Let $X = \bigcup S$, let $x \in X$ and $X' = X \setminus \{x\}$. Define ($\triangle$ denoting the symmetric difference of two sets)

$$S' = \{A \in S : A \triangle \{x\} \in S\},$$
$$S'' = \{A \in S : A \triangle \{x\} \notin S\}.$$

Clearly, $\sqcup$ denoting the disjoint union, $S = S' \sqcup S''$ and $S \cap X' = (S' \cap X') \sqcup (S'' \cap X')$. Moreover $|S'| = 2|S' \cap X'|$ and $|S''| = |S'' \cap X'|$. Thus $|S| = |S'| + |S''| = 2|S' \cap X'| + |S''| = |S \cap X'| + |S' \cap X'|$. Obviously $h(S \cap X') \leq h(S)$. Moreover $h(S' \cap X') = h(S') - 1$, because if $A \subset X'$ is shattered by S' (or equivalently by $S' \cap X'$), then $A \cup \{x\}$ is shattered by S' (we say that A is shattered by S when $S \cap A = \{0,1\}^A$). Using the induction hypothesis, we then see that $|S \cap X'| \leq \Phi_{|X'|}^{h(S)} + \Phi_{|X'|}^{h(S)-1}$. But as $|X'| = |X| - 1$, the righthand side of this inequality is equal to $\Phi_{|X|}^{h(S)}$, according to the recurrence equation satisfyied by Φ.

PROOF OF THEOREM 3.14: This is the well known Chernoff bound for the deviation of sums of Bernoulli r.v.: let $(\sigma_1, \dots, \sigma_n)$ be i.i.d. Bernoulli r.v. with parameter $1/2$. Let us notice that

$$\Phi_n^h = 2^n \mathbb{P}\left(\sum_{i=1}^n \sigma_i \leq h\right).$$

For any positive real number λ ,

$$\begin{aligned}\mathbb{P}(\sum_{i=1}^n \sigma_i \leq h) &\leq \exp(\lambda h)\mathbb{E}\left[\exp\left(-\lambda \sum_{i=1}^n \sigma_i\right)\right] \\ &= \exp\Bigl\{\lambda h + n\log\bigl\{\mathbb{E}\bigl[\exp\bigl(-\lambda\sigma_1\bigr)\bigr]\bigr\}\Bigr\}.\end{aligned}$$

Differentiating the right-hand side in λ shows that its minimal value is $\exp\bigl[-n\mathcal{K}(\frac{h}{n},\frac{1}{2})\bigr]$, where $\mathcal{K}(p,q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ is the Kullback divergence function between two Bernoulli distributions B_p and B_q of parameters p and q. Indeed the optimal value λ^* of λ is such that $h = n\frac{\mathbb{E}\bigl[\sigma_1\exp(-\lambda^*\sigma_1)\bigr]}{\mathbb{E}\bigl[\exp(-\lambda^*\sigma_1)\bigr]} = nB_{h/n}(\sigma_1)$. Therefore (using the fact that two Bernoulli distributions with the same expectations are equal)

$$\log\bigl\{\mathbb{E}\bigl[\exp(-\lambda^*\sigma_1)\bigr]\bigr\} = -\lambda^* B_{h/n}(\sigma_1) - \mathcal{K}(B_{h/n}, B_{1/2}) = -\lambda^*\tfrac{h}{n} - \mathcal{K}(\tfrac{h}{n},\tfrac{1}{2}).$$

The announced result then follows from the identity

$$\begin{aligned} H(p) &= \log(2) - \mathcal{K}(p,\tfrac{1}{2}) \\ &= p\log(p^{-1}) + (1-p)\log(1+\frac{p}{1-p}) \leq p\bigl[\log(p^{-1})+1\bigr]. \end{aligned}$$

3.2.3. VC dimension of linear rules with margin. The proof of the following theorem has been suggested to us by a similar proof presented in [18].

Theorem 3.15. *Consider a family of points $(x_1,\dots,x_n)$ in some Euclidean vector space E and a family of affine functions*

$$\mathcal{H} = \bigl\{g_{w,b} : E \to \mathbb{R}\,;\, w\in E, \|w\|=1, b\in\mathbb{R}\bigr\},$$

where

$$g_{w,b}(x) = \langle w,x\rangle - b, \qquad x\in E.$$

Assume that there is a set of thresholds $(b_i)_{i=1}^n \in \mathbb{R}^n$ such that for any $(y_i)_{i=1}^n \in \{-1,+1\}^n$, there is $g_{w,b}\in\mathcal{H}$ such that

$$\inf_{i=1}^{n}\bigl(g_{w,b}(x_i) - b_i\bigr)y_i \geq \gamma.$$

Let us also introduce the empirical variance of $(x_i)_{i=1}^n$,

$$\mathbb{V}\mathrm{ar}(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^n\Bigl\|x_i - \frac{1}{n}\sum_{j=1}^n x_j\Bigr\|^2.$$

In this case and with these notations,

$$\frac{\mathbb{V}\mathrm{ar}(x_1,\dots,x_n)}{\gamma^2} \geq \begin{cases} n-1 & \text{when } n \text{ is even,} \\ (n-1)\frac{n^2-1}{n^2} & \text{when } n \text{ is odd.}\end{cases} \tag{3.5}$$

Moreover, equality is reached when γ is optimal, $b_i = 0$, $i=1,\dots,n$ and $(x_1,\dots,x_n)$ is a regular simplex (i.e. when 2γ is the minimum distance between the convex hulls of any two subsets of $\{x_1,\dots,x_n\}$ and $\|x_i - x_j\|$ does not depend on $i\neq j$).

Proof. Let $(s_i)_{i=1}^n \in \mathbb{R}^n$ be such that $\sum_{i=1}^n s_i = 0$. Let σ be a uniformly distributed random variable with values in $\mathfrak{S}_n$, the set of permutations of the first n integers $\{1, \dots, n\}$. By assumption, for any value of σ, there is an affine function $g_{w,b} \in \mathcal{H}$ such that

$$\min_{i=1,\dots,n} \bigl[g_{w,b}(x_i) - b_i\bigr]\bigl[2\mathbb{1}(s_{\sigma(i)} > 0) - 1\bigr] \geq \gamma.$$

As a consequence

$$\begin{aligned}\left\langle \sum_{i=1}^n s_{\sigma(i)} x_i, w \right\rangle &= \sum_{i=1}^n s_{\sigma(i)} \bigl(\langle x_i, w\rangle - b - b_i\bigr) + \sum_{i=1}^n s_{\sigma(i)} b_i \\ &\geq \sum_{i=1}^n \gamma |s_{\sigma(i)}| + s_{\sigma(i)} b_i.\end{aligned}$$

Therefore, using the fact that the map $x \mapsto \Bigl(\max\bigl\{0, x\bigr\}\Bigr)^2$ is convex,

$$\begin{aligned}\mathbb{E}\left(\left\|\sum_{i=1}^n s_{\sigma(i)} x_i\right\|^2\right) &\geq \mathbb{E}\left[\left(\max\left\{0, \sum_{i=1}^n \gamma |s_{\sigma(i)}| + s_{\sigma(i)} b_i\right\}\right)^2\right] \\ &\geq \left(\max\left\{0, \sum_{i=1}^n \gamma \mathbb{E}\bigl(|s_{\sigma(i)}|\bigr) + \mathbb{E}\bigl(s_{\sigma(i)}\bigr) b_i\right\}\right)^2 = \gamma^2 \left(\sum_{i=1}^n |s_i|\right)^2,\end{aligned}$$

where $\mathbb{E}$ is the expectation with respect to the random permutation σ. On the other hand

$$\mathbb{E}\left(\left\|\sum_{i=1}^n s_{\sigma(i)} x_i\right\|^2\right) = \sum_{i=1}^n \mathbb{E}(s_{\sigma(i)}^2) \|x_i\|^2 + \sum_{i \neq j} \mathbb{E}(s_{\sigma(i)} s_{\sigma(j)}) \langle x_i, x_j\rangle.$$

Moreover

$$\mathbb{E}(s_{\sigma(i)}^2) = \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n s_{\sigma(i)}^2\right) = \frac{1}{n} \sum_{i=1}^n s_i^2.$$

In the same way, for any $i \neq j$,

$$\begin{aligned}\mathbb{E}\left(s_{\sigma(i)} s_{\sigma(j)}\right) &= \frac{1}{n(n-1)} \mathbb{E}\left(\sum_{i \neq j} s_{\sigma(i)} s_{\sigma(j)}\right) \\ &= \frac{1}{n(n-1)} \sum_{i \neq j} s_i s_j\end{aligned}$$

$$= \frac{1}{n(n-1)} \left[\underbrace{\left(\sum_{i=1}^n s_i \right)^2}_{=0} - \sum_{i=1}^n s_i^2 \right]$$

$$= - \frac{1}{n(n-1)} \sum_{i=1}^n s_i^2.$$

Thus

$$\mathbb{E} \left(\left\| \sum_{i=1}^n s_{\sigma(i)} x_i \right\|^2 \right) = \left(\sum_{i=1}^n s_i^2 \right) \left[\frac{1}{n} \sum_{i=1}^n \lVert x_i \rVert^2 - \frac{1}{n(n-1)} \sum_{i \neq j} \langle x_i, x_j \rangle \right]$$

$$= \left(\sum_{i=1}^n s_i^2 \right) \left[\left(\frac{1}{n} + \frac{1}{n(n-1)} \right) \sum_{i=1}^n \lVert x_i \rVert^2 \right.$$

$$\left. - \frac{1}{n(n-1)} \left\| \sum_{i=1}^n x_i \right\|^2 \right]$$

$$= \frac{n}{n-1} \left(\sum_{i=1}^n s_i^2 \right) \mathbb{V}\mathrm{ar}(x_1, \dots, x_n).$$

We have proved that

$$\frac{\mathbb{V}\mathrm{ar}(x_1, \dots, x_n)}{\gamma^2} \geq \frac{(n-1) \left(\displaystyle\sum_{i=1}^n \lvert s_i \rvert \right)^2}{n \displaystyle\sum_{i=1}^n s_i^2}.$$

This can be used with $s_i = \mathbb{1}(i \leq \frac{n}{2}) - \mathbb{1}(i > \frac{n}{2})$ in the case when n is even and $s_i = \frac{2}{(n-1)} \mathbb{1}(i \leq \frac{n-1}{2}) - \frac{2}{n+1} \mathbb{1}(i > \frac{n-1}{2})$ in the case when n is odd to establish the first inequality (3.5) of the theorem.

Checking that equality is reached for the simplex is an easy computation when the simplex $(x_i)_{i=1}^n \in (\mathbb{R}^n)^n$ is parametrized in such a way that

$$x_i(j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed the distance between the convex hulls of any two subsets of the simplex is the distance between their mean values (i.e. centers of mass). $\square$

3.2.4. Application to Support Vector Machines. We are going to apply Theorem 3.15 (page 139) to Support Vector Machines in the transductive case. So let us consider $(X_i, Y_i)_{i=1}^{(k+1)N}$ distributed according to some partially exchangeable distribution $\mathbb{P}$ and assume that $(X_i)_{i=1}^{(k+1)N}$ and $(Y_i)_{i=1}^{N}$ are observed. Let us consider some positive kernel K on $\mathcal{X}$. For any K-separable training set of the form $Z' = (X_i, y'_i)_{i=1}^{(k+1)N}$, where $(y'_i)_{i=1}^{(k+1)N} \in \mathcal{Y}^{(k+1)N}$, let $\hat{f}_{Z'}$ be the Support Vector Machine defined by K and Z' and let $\gamma(Z')$ be its margin. Let

$$R^2 = \max_{i=1,\dots,(k+1)N} K(X_i, X_i) + \frac{1}{(k+1)^2 N^2} \sum_{j=1}^{(k+1)N} \sum_{k=1}^{(k+1)N} K(X_j, X_k) \\ - \frac{2}{(k+1)N} \sum_{j=1}^{(k+1)N} K(X_i, X_j).$$

(This is an easily computable upper-bound for the radius of some ball containing the image of $(X_1, \dots, X_{(k+1)N})$ in feature space.)

Let us define for any integer h the margins

$$\gamma_{2h} = (2h-1)^{-1/2} \text{ and } \gamma_{2h+1} = \left[2h \left(1 - \frac{1}{(2h+1)^2} \right) \right]^{-1/2}. \tag{3.6}$$

Let us consider for any $h = 1, \dots, N$ the exchangeable model

$$\mathcal{R}_h = \bigl\{ \hat{f}_{Z'} : Z' = (X_i, y'_i)_{i=1}^{(k+1)N} \text{ is } K\text{-separable and } \gamma(Z') \geq R\gamma_h \bigr\}.$$

The family of models $\mathcal{R}_h$, $h = 1, \dots, N$ is nested, and we know from Theorem 3.15 (page 139) and Theorems 3.13 (page 138) and 3.14 (page 138) that

$$\log\bigl(|\mathcal{R}_h|\bigr) \leq h \log\bigl(\tfrac{(k+1)eN}{h}\bigr).$$

We can then consider on the large model $\mathcal{R} = \bigsqcup_{h=1}^{N} \mathcal{R}_h$ (the disjoint union of the submodels) an exchangeable prior π which is uniform on each $\mathcal{R}_h$ and is such that $\pi(\mathcal{R}_h) \geq \frac{1}{h(h+1)}$. Applying Theorem 2.8 (page 104) we get

Proposition 3.16. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any $h = 1, \dots, N$, any Support Vector Machine $f \in \mathcal{R}_h$,*

$$r_2(f) \leq \\ \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{1 - \exp\left[-\frac{\lambda}{N} r_1(f) - \frac{h}{N} \log\left(\frac{e(k+1)N}{h} \right) - \frac{\log[h(h+1)] - \log(\epsilon)}{N} \right]}{1 - \exp(-\frac{\lambda}{N})} \\ - \frac{r_1(f)}{k}.$$

Searching the whole model $\mathcal{R}_h$ may be unfeasible, nonetheless any heuristic can be applied to choose f. For instance, a Support Vector Machine f' can be trained from the training set $(X_i, Y_i)_{i=1}^N$ and then $(y'_i)_{i=1}^{(k+1)N}$ can be set to $y'_i = \mathrm{sign}(f'(X_i))$, $i = 1, \dots, (k+1)N$.

3.2.5. Inductive margin bounds for Support Vector Machines. In order to establish inductive margin bounds, we will need a different combinatorial lemma. It is due to [1]. We will reproduce their proof with some tiny improvements on the values of constants.

Let us consider the finite case when $\mathcal{X} = \{1, \dots, n\}$, $\mathcal{Y} = \{1, \dots, b\}$ and $b \geq 3$ (the question we will study would be meaningless in the case when $b \leq 2$). Assume as usual that we are dealing with a prescribed set of classification rules
$\mathcal{R} = \bigl\{ f : \mathcal{X} \to \mathcal{Y} \bigr\}$. Let us say that a pair (A, s), where $A \subset \mathcal{X}$ is a non empty set of shapes and $s : A \to \{2, \dots, b-1\}$ a threshold function, is *shattered* by the set of functions $F \subset \mathcal{R}$ if for any $(\sigma_x)_{x \in A} \in \{-1, +1\}^A$, there exists some $f \in F$ such that $\min_{x \in A} \sigma_x \bigl[f(x) - s(x) \bigr] \geq 1$.

DEFINITION 3.5. Let the *fat shattering dimension* of $(\mathcal{X}, \mathcal{R})$ be the maximal size $|A|$ of the first component of the pairs which are shattered by $\mathcal{R}$.

Let us say that a subset of classification rules $F \subset \mathcal{Y}^{\mathcal{X}}$ is *separated* whenever for any pair $(f, g) \in F^2$ such that $f \neq g$, $\|f - g\|_\infty = \max_{x \in \mathcal{X}} |f(x) - g(x)| \geq 2$. Let $\mathfrak{M}(\mathcal{R})$ be the maximum size $|F|$ of separated subsets F of $\mathcal{R}$. Note that if F is a separated subset of $\mathcal{R}$ such that $|F| = \mathfrak{M}(\mathcal{R})$, then it is a 1-net for the $\mathcal{L}_\infty$ distance: for any function $f \in \mathcal{R}$ there exists $g \in F$ such that $\|f - g\|_\infty \leq 1$ (otherwise f could be added to F to create a larger separated set).

LEMMA 3.17. *With the above notations, whenever the fat shattering dimension of* $(\mathcal{X}, \mathcal{R})$ *is not greater than* h,

$$\log\bigl[\mathfrak{M}(\mathcal{R})\bigr] < \log\bigl[(b-1)(b-2)n\bigr] \left\{ \frac{\log\bigl[\sum_{i=1}^h \binom{n}{i} (b-2)^i\bigr]}{\log(2)} + 1 \right\} + \log(2)$$
$$\leq \log\bigl[(b-1)(b-2)n\bigr] \left\{ \left[\log\Bigl[\tfrac{(b-2)n}{h}\Bigr] + 1 \right] \frac{h}{\log(2)} + 1 \right\} + \log(2).$$

PROOF. For any set of functions $F \subset \mathcal{Y}^{\mathcal{X}}$, let $t(F)$ be the number of pairs (A, s) shattered by F. Let $t(m, n)$ be the minimum of $t(F)$ over all *separated* sets of functions $F \subset \mathcal{Y}^{\mathcal{X}}$ of size $|F| = m$ (n is here to recall that the shape

space $\mathcal{X}$ is made of n shapes). For any m such that $t(m,n) > \sum_{i=1}^{h} \binom{n}{i}(b-2)^i$, it is clear that any separated set of functions of size $|F| \geq m$ shatters at least one pair (A,s) such that $|A| > h$. Indeed, $t(m,n)$ is clearly from its definition a non decreasing function of m, so that $t(|F|,n) > \sum_{i=1}^{h} \binom{n}{i}(b-2)^i$. Moreover there are only $\sum_{i=1}^{h} \binom{n}{i}(b-2)^i$ pairs (A,s) such that $|A| \leq h$. As a consequence, whenever the fat shattering dimension of $(\mathcal{X},\mathcal{R})$ is not greater than h we have $\mathfrak{M}(\mathcal{R}) < m$.

It is clear that for any $n \geq 1$, $t(2,n) = 1$.

Lemma 3.18. *For any $m \geq 1$, $t\bigl[mn(b-1)(b-2),n\bigr] \geq 2t\bigl[m,n-1\bigr]$, and therefore $t\bigl[2n(n-1)\dots(n-r+1)(b-1)^r(b-2)^r,n\bigr] \geq 2^r$.*

Proof. Let $F = \{f_1,\dots,f_{mn(b-1)(b-2)}\}$ be some separated set of functions of size $mn(b-1)(b-2)$. For any pair (f_{2i-1},f_{2i}), $i = 1,\dots,mn(b-1)(b-2)/2$, there is $x_i \in \mathcal{X}$ such that $|f_{2i-1}(x_i) - f_{2i}(x_i)| \geq 2$. Since $|\mathcal{X}| = n$, there is $x \in \mathcal{X}$ such that $\sum_{i=1}^{mn(b-1)(b-2)/2} \mathbb{1}(x_i = x) \geq m(b-1)(b-2)/2$. Let $I = \{i : x_i = x\}$. Since there are $(b-1)(b-2)/2$ pairs $(y_1,y_2) \in \mathcal{Y}^2$ such that $1 \leq y_1 < y_2 - 1 \leq b-1$, there is some pair (y_1,y_2), such that $1 \leq y_1 < y_2 \leq b$ and such that $\sum_{i\in I} \mathbb{1}(\{y_1,y_2\} = \{f_{2i-1}(x),f_{2i}(x)\}) \geq m$. Let $J = \bigl\{i \in I : \{f_{2i-1}(x),f_{2i}(x)\} = \{y_1,y_2\}\bigr\}$. Let

$$
\begin{aligned}
F_1 &= \{f_{2i-1} : i \in J, f_{2i-1}(x) = y_1\} \cup \{f_{2i} : i \in J, f_{2i}(x) = y_1\},\\
F_2 &= \{f_{2i-1} : i \in J, f_{2i-1}(x) = y_2\} \cup \{f_{2i} : i \in J, f_{2i}(x) = y_2\}.
\end{aligned}
$$

Obviously $|F_1| = |F_2| = |J| = m$. Moreover the restrictions of the functions of F_1 to $\mathcal{X} \setminus \{x\}$ are separated, and it is the same with F_2. Thus F_1 strongly shatters at least $t(m,n-1)$ pairs (A,s) such that $A \subset \mathcal{X} \setminus \{x\}$ and it is the same with F_2. Eventually, if the pair (A,s) where $A \subset \mathcal{X} \setminus \{x\}$ is both shattered by F_1 and F_2, then $F_1 \cup F_2$ shatters also $(A \cup \{x\}, s')$ where $s'(x') = s(x')$ for any $x' \in A$ and $s'(x) = \lfloor \frac{y_1+y_2}{2} \rfloor$. Thus $F_1 \cup F_2$, and therefore F, shatters at least $2t(m,n-1)$ pairs (A,s). $\square$

Resuming the proof of lemma 3.17, let us choose for r the smallest integer such that $2^r > \sum_{i=1}^{h} \binom{n}{i}(b-2)^i$, which is no greater than

$$
\left\{ \frac{\log\bigl[\sum_{i=1}^{h}\binom{n}{i}(b-2)^i\bigr]}{\log(2)} + 1 \right\}.
$$

In the case when $1 \leq n \leq r$,

$$
\log(\mathfrak{M}(\mathcal{R})) < |\mathcal{X}|\log(|\mathcal{Y}|) = n\log(b) \leq r\log(b) \leq r\log\bigl[(b-1)(b-2)n\bigr] + \log(2),
$$

which proves the lemma. In the remaining case $n > r$,

$$
\begin{aligned}
t\bigl[2n^r(b-1)^r(b-2)^r, n\bigr] \\
&\geq t\bigl[2n(n-1)\dots(n-r+1)(b-1)^r(b-2)^r, n\bigr] \\
&\qquad > \sum_{i=1}^{h} \binom{n}{i} (b-2)^i.
\end{aligned}
$$

Thus $|\mathfrak{M}(\mathcal{R})| < 2\Bigl[(b-2)(b-1)n\Bigr]^r$ as claimed. □

In order to apply this combinatorial lemma to Support Vector Machines, let us consider now the case of separating hyperplanes in $\mathbb{R}^d$ (the generalization to Support Vector Machines being straightforward). Assume that $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. For any sample $(X)_{i=1}^{(k+1)N}$, let

$$
R(X_1^{(k+1)N}) = \max\{\|X_i\| \, : 1 \leq i \leq (k+1)N\}.
$$

Let us consider the set of parameters

$$
\Theta = \bigl\{(w, b) \in \mathbb{R}^d \times \mathbb{R} \, : \|w\| = 1\bigr\}.
$$

For any $(w, b) \in \Theta$, let $g_{w,b}(x) = \langle w, x\rangle - b$. Let h be some fixed integer and let $\gamma = R(X_1^{(k+1)N})\gamma_h$, where γ_h is defined by equation (3.6) on page 142.

Let us define $\zeta : \mathbb{R} \to \mathbb{Z}$ by

$$
\zeta(r) = \begin{cases}
-5 & \text{when} \quad r \leq -4\gamma, \\
-3 & \text{when} \quad -4\gamma < r \leq -2\gamma, \\
-1 & \text{when} \quad -2\gamma < r \leq 0, \\
+1 & \text{when} \quad 0 < r \leq 2\gamma, \\
+3 & \text{when} \quad 2\gamma < r \leq 4\gamma, \\
+5 & \text{when} \quad 4\gamma < r.
\end{cases}
$$

Let $G_{w,b}(x) = \zeta\bigl[g_{w,b}(x)\bigr]$. The fat shattering dimension (as defined in 3.5) of

$$
\Bigl(X_1^{(k+1)N}, \bigl\{(G_{w,b} + 7)/2 : (w, b) \in \Theta\bigr\}\Bigr)
$$

is not greater than h (according to Theorem 3.15, page 139), therefore there is some set $\mathcal{F}$ of functions from $X_1^{(k+1)N}$ to $\{-5, -3, -1, +1, +3, +5\}$ such that

$$
\log\bigl(|\mathcal{F}|\bigr) \leq \log\bigl[20(k+1)N\bigr] \left\{ \frac{h}{\log(2)} \left[\log\left(\frac{4(k+1)N}{h} \right) + 1 \right] + 1 \right\} + \log(2).
$$

and for any $(w,b) \in \Theta$, there is $f_{w,b} \in \mathcal{F}$ such that $\sup\bigl\{ \lvert f_{w,b}(X_i) - G_{w,b}(X_i) \rvert : i = 1, \dots, (k+1)N \bigr\} \leq 2$. Moreover, the choice of $f_{w,b}$ may be required to depend on $(X_i)_{i=1}^{(k+1)N}$ in an exchangeable way. Similarly to Theorem 2.8 (page 104), it can be proved that for any partially exchangeable probability distribution $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $f_{w,b} \in \mathcal{F}$,

$$
\begin{aligned}
\frac{1}{kN} & \sum_{i=N+1}^{(k+1)N} \mathbb{1}\bigl[f_{w,b}(X_i) Y_i \leq 1 \bigr] \\
& \leq \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \bigl[1 - \exp(-\tfrac{\lambda}{N}) \bigr]^{-1} \biggl\{ 1 - \\
& \qquad \exp \biggl[- \frac{\lambda}{N^2} \sum_{i=1}^N \mathbb{1}\bigl[f_{w,b}(X_i) Y_i \leq 1 \bigr] - \frac{\log\bigl(\lvert \mathcal{F} \rvert\bigr) - \log(\epsilon)}{N} \biggr] \biggr\} \\
& \qquad\qquad - \frac{1}{kN} \sum_{i=1}^N \mathbb{1}\bigl[f_{w,b}(X_i) Y_i \leq 1 \bigr].
\end{aligned}
$$

Let us remark that

$$
\mathbb{1}\Bigl\{ 2\mathbb{1}\bigl[g_{w,b}(X_i) \geq 0 \bigr] - 1 \neq Y_i \Bigr\} = \mathbb{1}\bigl[G_{w,b}(X_i) Y_i < 0 \bigr] \leq \mathbb{1}\bigl[f_{w,b}(X_i) Y_i \leq 1 \bigr]
$$

and

$$
\mathbb{1}\bigl[f_{w,b}(X_i) Y_i \leq 1 \bigr] \leq \mathbb{1}\bigl[G_{w,b}(X_i) Y_i \leq 3 \bigr] \leq \mathbb{1}\bigl[g_{w,b}(X_i) Y_i \leq 4\gamma \bigr].
$$

This proves the following theorem.

Theorem 3.19. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any $(w,b) \in \Theta$,*

$$
\begin{aligned}
\frac{1}{kN} & \sum_{i=N+1}^{(k+1)N} \mathbb{1}\Bigl\{ 2\mathbb{1}\bigl[g_{w,b}(X_i) \geq 0 \bigr] - 1 \neq Y_i \Bigr\} \\
& \leq \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+, h \in \mathbb{N}^*} \bigl[1 - \exp(-\tfrac{\lambda}{N}) \bigr]^{-1} \biggl\{ 1 - \\
& \exp \biggl[- \frac{\lambda}{N^2} \sum_{i=1}^N \mathbb{1}\bigl[g_{w,b}(X_i) Y_i \leq 4R\gamma_h \bigr] \\
& \quad - \frac{\log\bigl[20(k+1)N\bigr] \Bigl\{ \frac{h}{\log(2)} \log\Bigl(\frac{4e(k+1)N}{h} \Bigr) + 1 \Bigr\} + \log\Bigl[\frac{2h(h+1)}{\epsilon} \Bigr]}{N} \biggr] \biggr\}
\end{aligned}
$$

$$- \frac{1}{kN} \sum_{i=1}^{N} \mathbb{1} \bigl[g_{w,b}(X_i) Y_i \leq 4 R \gamma_h \bigr].$$

As a consequence, we obtain with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $(w,b) \in \Theta$ such that

$$\gamma = \min_{i=1,\dots,N} g_{w,b}(X_i) Y_i > 0,$$

$$\begin{aligned} \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} & \mathbb{1} \bigl[g_{w,b}(X_i) Y_i < 0 \bigr] \\ & \leq \tfrac{k+1}{k} \biggl\{ 1 - \exp \biggl[- \tfrac{\log \bigl[20(k+1)N \bigr]}{N} \Bigl\{ \tfrac{16R^2 + 2\gamma^2}{\log(2)\gamma^2} \log \Bigl(\tfrac{e(k+1)N\gamma^2}{4R^2} \Bigr) + 1 \Bigr\} \\ & \qquad \qquad + \frac{1}{N} \log(\tfrac{\epsilon}{2}) \biggr] \biggr\}. \end{aligned}$$

This inequality compares favourably with similar inequalities in [18], which moreover do not extend to the margin quantile case as this one.

Let us also remark that it is easy to circonvent the fact that R is not observed when the test set $X_{N+1}^{(k+1)N}$ is not observed.

Indeed, we can consider the sample obtained by projecting $X_1^{(k+1)N}$ on some ball of fixed radius $R_{\max}$, putting

$$t_{R_{\max}}(X_i) = \min \left\{ 1, \frac{R_{\max}}{\lVert X_i \rVert} \right\} X_i.$$

We can further consider an atomic prior distribution $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$ bearing on $R_{\max}$, to obtain a uniform result through a union bound. As a consequence of the previous theorem indeed,

Corollary 3.20. *For any atomic prior $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$, for any partially exchangeable probability measure $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, with $\mathbb{P}$ probability at least $1 - \epsilon$, for any $(w,b) \in \Theta$, any $R_{\max} \in \mathbb{R}_+$,*

$$
\begin{aligned}
&\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} \mathbb{1}\Bigl\{2\mathbb{1}\bigl[g_{w,b}\circ t_{R_{\max}}(X_i)\geq 0\bigr]-1\neq Y_i\Bigr\}\\
&\leq \frac{k+1}{k}\inf_{\lambda\in\mathbb{R}_+,h\in\mathbb{N}^*}\bigl[1-\exp(-\tfrac{\lambda}{N})\bigr]^{-1}\Biggl\{1-\\
&\quad\exp\Biggl[-\frac{\lambda}{N^2}\sum_{i=1}^{N}\mathbb{1}\bigl[g_{w,b}\circ t_{R_{\max}}(X_i)Y_i\leq 4R_{\max}\gamma_h\bigr]\\
&\qquad -\frac{\log\bigl[20(k+1)N\bigr]\Bigl\{\frac{h}{\log(2)}\log\Bigl(\frac{4e(k+1)N}{h}\Bigr)+1\Bigr\}+\log\Bigl[\frac{2h(h+1)}{\epsilon\nu(R_{\max})}\Bigr]}{N}\Biggr]\Biggr\}\\
&\qquad\qquad -\frac{1}{kN}\sum_{i=1}^{N}\mathbb{1}\bigl[g_{w,b}\circ t_{R_{\max}}(X_i)Y_i\leq 4R_{\max}\gamma_h\bigr].
\end{aligned}
$$

4. Appendix: classification by thresholding

In this appendix, we show how the bounds given in the first section of this monograph can be computed in practice on a simple example: the case when the classification is performed by comparing a series of measurements to threshold values. Let us mention that our description covers the case when the same measurement is compared to several thresholds, since it is enough to repeat a measurement in the list of measurements describing a pattern to cover this case.

4.1. Description of the model.

Let us assume that the patterns we want to classify are described through h real valued measurements normalized in the range $(0,1)$. In this setting the pattern space can thus be defined as $\mathcal{X}=(0,1)^h$.

Consider the threshold set $\mathcal{T}=(0,1)^h$ and the response set $\mathcal{R}=\mathcal{Y}^{\{0,1\}^h}$. For any $t\in(0,1)^h$ and any $a:\{0,1\}^h\rightarrow\mathcal{Y}$, let

$$
f_{(t,a)}(x)=a\Bigl\{\bigl[\mathbb{1}(x^j\geq t_j)\bigr]_{j=1}^h\Bigr\},\quad x\in\mathcal{X},
$$

where x^j is the jth coordinate of $x\in\mathcal{X}$. Thus our parameter set here is $\Theta=\mathcal{T}\times\mathcal{R}$. Let us consider on $\mathcal{T}$ the Lebesgue measure L and on $\mathcal{R}$ the uniform probability distribution U. Let our prior distribution be $\pi=L\otimes U$. Let us define for any threshold sequence $t\in\mathcal{T}$

$$
\Delta_t=\Bigl\{t'\in\mathcal{T}:\overline{(t'_j,t_j)}\cap\{X_i^j;i=1,\dots,N\}=\varnothing, j=1,\dots,h\Bigr\},
$$

where X_i^j is the jth coordinate of the sample pattern X_i, and where the interval $\overline{(t'_j, t_j)}$ of the real line is defined as the convex hull of the two point set $\{t'_j, t_j\}$, whether $t'_j \leq t_j$ or not. We see that Δ_t is the set of thresholds giving the same response as t on the training patterns. Let us consider for any $t \in \mathcal{T}$ the middle

$$m(\Delta_t) = \frac{\int_{\Delta_t} t' L(dt')}{L(\Delta_t)}$$

of Δ_t. The set Δ_t being a product of intervals, its middle is the point whose coordinates are the middle of these intervals. Let us introduce the finite set T composed of the middles of the cells Δ_t, which can be defined as

$$T = \{t \in \mathcal{T} : t = m(\Delta_t)\}.$$

It is easy to see that $|T| \leq (N+1)^h$ and that $|\mathcal{R}| = |\mathcal{Y}|^{2^h}$.

4.2. Computation of inductive bounds. For any parameter $(t, a) \in \mathcal{T} \times \mathcal{R} = \Theta$, let us consider the posterior distribution defined by its density

$$\frac{d\rho_{(t,a)}}{d\pi}(t', a') = \frac{\mathbb{1}\bigl(t' \in \Delta_t\bigr)\mathbb{1}\bigl(a' = a\bigr)}{\pi\bigl(\Delta_t \times \{a\}\bigr)}.$$

Let us notice that we are in fact considering a finite number of posterior distributions, since $\rho_{(t,a)} = \rho_{(m(\Delta_t),a)}$, where $m(\Delta_t) \in T$. Let us also mention that for any exchangeable sample distribution $\mathbb{P} \in \mathcal{M}_+^1\bigl[(\mathcal{X} \times \mathcal{Y})^{N+1}\bigr]$ and any thresholds $t \in \mathcal{T}$,

$$\mathbb{P}\Bigl[\overline{(X_{N+1}^j, t_j)} \cap \{X_i^j, i = 1, \dots, N\} = \varnothing\Bigr] \leq \frac{2}{N+1}.$$

Thus, for any $(t, a) \in \Theta$,

$$\mathbb{P}\Bigl\{\rho_{(t,a)}\bigl[f_{\cdot}(X_{N+1})\bigr] \neq f_{(t,a)}(X_{N+1})\Bigr\} \leq \frac{2h}{N+1},$$

showing that the classification produced by $\rho_{(t,a)}$ on new examples is most of the time non random (this result is only indicative, since it is concerned with a non random choice of (t, a)).

Let us then compute the various quantities needed to apply the results of the first section, focussing our attention of Theorem 1.39 (page 63):

It is to be noted first of all that $\rho_{(t,a)}(r) = r[(t, a)]$. The entropy term is such that

$$\mathcal{K}(\rho_{t,a}, \pi) = -\log\bigl[\pi\bigl(\Delta_t \times \{r\}\bigr)\bigr] = -\log\bigl[L(\Delta_t)\bigr] + 2^h \log\bigl(|\mathcal{Y}|\bigr).$$

Let us notice accordingly that

$$\min_{(t,a)\in\Theta} \mathcal{K}(\rho_{(t,a)}, \pi) \leq h\log(N+1) + 2^h \log\bigl(|\mathcal{Y}|\bigr).$$

Let us introduce the counters

$$b_y^t(c) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\Bigl\{Y_i = y \text{ and } \bigl[\mathbb{1}(X_i^j \geq t_j)\bigr]_{j=1}^h = c\Bigr\}, \qquad t \in T, c \in \{0,1\}^h, y \in \mathcal{Y},$$

$$b^t(c) = \sum_{y\in\mathcal{Y}} b_y^t(c) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\Bigl\{\bigl[\mathbb{1}(X_i^j \geq t_j)\bigr]_{j=1}^h = c\Bigr\}, \qquad t \in T, c \in \{0,1\}^h.$$

Since

$$r[(t,a)] = \sum_{c\in\{0,1\}^h} \bigl[b^t(c) - b_{a(c)}^t(c)\bigr],$$

the partition function of the Gibbs estimator can be computed as

$$\begin{aligned}
\pi\bigl[\exp(-\lambda r)\bigr] &= \sum_{t\in T} L(\Delta_t) \sum_{a\in\mathcal{R}} \frac{1}{|\mathcal{Y}|^{2^h}} \exp\Biggl[-\lambda \sum_{i=1}^N \mathbb{1}\bigl[Y_i \neq f_{(t,a)}(X_i)\bigr]\Biggr] \\
&= \sum_{t\in T} L(\Delta_t) \sum_{a\in\mathcal{R}} \frac{1}{|\mathcal{Y}|^{2^h}} \exp\Biggl[-\lambda \sum_{c\in\{0,1\}^h} \bigl[b^t(c) - b_{a(c)}^t(c)\bigr]\Biggr] \\
&= \sum_{t\in T} L(\Delta_t) \prod_{c\in\{0,1\}^h} \Biggl[\frac{1}{|\mathcal{Y}|} \sum_{y\in\mathcal{Y}} \exp\Bigl(-\lambda\bigl[b^t(c) - b_y^t(c)\bigr]\Bigr)\Biggr].
\end{aligned}$$

We see that the number of operations needed to compute $\pi\bigl[\exp(-\lambda r)\bigr]$ is proportional to $|T| \times 2^h \times |\mathcal{Y}| \leq (N+1)^h 2^h |\mathcal{Y}|$. An exact computation will therefore be feasible only for small values of N and h. For higher values, a Monte Carlo approximation of this sum will have to be performed instead.

If we want to compute the bound provided by Theorem 1.39 (page 63), we need also to compute, for any fixed parameter $\theta \in \Theta$, quantities of the type

$$\pi_{\exp(-\lambda r)}\Bigl\{\exp\bigl[\xi m'(\cdot,\theta)\bigr]\Bigr\} = \pi_{\exp(-\lambda r)}\Bigl\{\exp\bigl[\xi \rho_\theta(m')\bigr]\Bigr\}, \quad \lambda, \xi \in \mathbb{R}_+.$$

To this purpose we need to introduce

$$\overline{b}_y^t(\theta, c) = \frac{1}{N}\sum_{i=1}^N \Bigl|\mathbb{1}\bigl[f_\theta(X_i) \neq Y_i\bigr] - \mathbb{1}(y \neq Y_i)\Bigr| \mathbb{1}\bigl\{\bigl[\mathbb{1}(X_i^j \geq t_j)\bigr]_{j=1}^h = c\bigr\}.$$

Similarly to what has been done previously, we obtain

$$\pi\Bigl\{\exp\bigl[-\lambda r + \xi m'(\cdot,\theta)\bigr]\Bigr\} \\ = \sum_{t \in T} L(\Delta_t) \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda\bigl[b^t(c) - b^t_y(c)\bigr] + \xi \overline{b}^t_y(\theta, c)\Bigr)\right].$$

We can then compute

$$\begin{aligned} \pi_{\exp(-\lambda r)}(r) &= -\frac{\partial}{\partial \lambda} \log\Bigl\{\pi\bigl[\exp(-\lambda r)\bigr]\Bigr\}, \\ \pi_{\exp(-\lambda r)}\Bigl\{\exp\bigl[\xi \rho_\theta(m')\bigr]\Bigr\} &= \frac{\pi\Bigl\{\exp\bigl[-\lambda r + \xi m'(\cdot,\theta)\bigr]\Bigr\}}{\pi\bigl[\exp(-\lambda r)\bigr]}, \\ \pi_{\exp(-\lambda r)}\bigl[m'(\cdot,\theta)\bigr] &= \frac{\partial}{\partial \xi}_{|\xi=0} \log\Bigl[\pi\Bigl\{\exp\bigl[-\lambda r + \xi m'(\cdot,\theta)\bigr]\Bigr\}\Bigr]. \end{aligned}$$

This is all we need to compute $B(\rho_\theta, \beta, \gamma)$ (and also $B(\pi_{\exp(-\lambda r)}, \beta, \gamma)$) in Theorem 1.39 (page 63), using the approximation

$$\log\Bigl\{\pi_{\exp(-\lambda_1 r)}\Bigl[\exp\bigl\{\xi \pi_{\exp(-\lambda_2 r)}(m')\bigr\}\Bigr]\Bigr\} \\ \leq \log\Bigl\{\pi_{\exp(-\lambda_1 r)}\Bigl[\exp\bigl\{\xi m'(\cdot,\theta)\bigr\}\Bigr]\Bigr\} + \xi \pi_{\exp(-\lambda_2 r)}\bigl[m'(\cdot,\theta)\bigr], \quad \xi \geq 0.$$

Let us also explain how to apply the posterior distribution $\rho_{(t,a)}$, in other words our randomized estimated classification rule, to a new pattern X_{N+1}:

$$\rho_{(t,a)}\bigl[f_\cdot(X_{N+1}) = y\bigr] = L(\Delta_t)^{-1} \int_{\Delta_t} \mathbb{1}\Bigl[a\bigl\{\bigl[\mathbb{1}(X^j_{N+1} \geq t'_j)\bigr]_{j=1}^h\bigr\} = y\Bigr] L(dt') \\ = L(\Delta_t)^{-1} \sum_{c \in \{0,1\}^h} L\Bigl(\Bigl\{t' \in \Delta_t : \bigl[\mathbb{1}(X^j_{N+1} \geq t'_j)\bigr]_{j=1}^h = c\Bigr\}\Bigr) \mathbb{1}\bigl[a(c) = y\bigr].$$

Let us define for short

$$\Delta_t(c) = \Bigl\{t' \in \Delta_t : \bigl[\mathbb{1}(X^j_{N+1} \geq t'_j)\bigr]_{j=1}^h = c\Bigr\}, \qquad c \in \{0,1\}^h.$$

With this notation

$$\rho_{(t,a)}\bigl[f_\cdot(X_{N+1}) = y\bigr] = L\bigl(\Delta_t\bigr)^{-1} \sum_{c \in \{0,1\}^h} L\bigl[\Delta_t(c)\bigr] \mathbb{1}\bigl[a(c) = y\bigr].$$

We can compute in the same way the probabilities for the label of the new pattern under the Gibbs posterior distribution:

$$\pi_{\exp(-\lambda r)}\bigl[f.(X_{N+1}) = y'\bigr]$$
$$= \Biggl\{ \sum_{t \in T} \prod_{c \in \{0,1\}^h} \biggl[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda \bigl[b^t(c) - b^t_y(c)\bigr]\Bigr)\biggr]$$
$$\times \sum_{c \in \{0,1\}^h} L\bigl[\Delta_t(c)\bigr] \frac{\sum_{y \in \mathcal{Y}} \mathbb{1}(y = y') \exp\bigl\{-\lambda\bigl[b^t(c) - b^t_y(c)\bigr]\bigr\}}{\sum_{y \in \mathcal{Y}} \exp\bigl\{-\lambda\bigl[b^t(x) - b^t_y(c)\bigr]\bigr\}} \Biggr\}$$
$$\times \Biggl\{ \sum_{t \in T} L(\Delta_t) \prod_{c \in \{0,1\}^h} \biggl[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda\bigl[b^t(c) - b^t_y(c)\bigr]\Bigr)\biggr]\Biggr\}^{-1}.$$

4.3. Transductive bounds. In the case when we observe the patterns of a shadow sample $(X_i)_{i=N+1}^{(k+1)N}$ on top of the training sample $(X_i, Y_i)_{i=1}^N$, we can introduce the set of thresholds responding as t on the extended sample $(X_i)_{i=1}^{(k+1)N}$

$$\overline{\Delta}_t = \Bigl\{ t' \in \mathcal{T} : \overline{(t'_j, t_j)} \cap \bigl\{X_i^j ; i = 1, \dots, (k+1)N\bigr\} = \varnothing, j = 1, \dots, h \Bigr\},$$

consider the set

$$\overline{T} = \bigl\{ t \in \mathcal{T} : t = m(\overline{\Delta}_t)\bigr\},$$

of the middle points of the cells $\overline{\Delta}_t$, $t \in \mathcal{T}$, and replace the Lebesgue measure $L \in \mathcal{M}_+^1\bigl[(0,1)^h\bigr]$ of the previous section with the uniform probability measure $\overline{L}$ on $\overline{T}$. We can then consider $\pi = \overline{L} \otimes U$, where U is as previously the uniform probability measure on $\mathcal{R}$. This gives obviously an exchangeable posterior distribution and therefore qualifies π for transductive bounds. Let us notice that $|\overline{T}| \leq \bigl[(k+1)N + 1\bigr]^h$, and therefore that $\pi(t,a) \geq \bigl[(k+1)N+1\bigr]^{-h} |\mathcal{Y}|^{-2^h}$, for any $(t,a) \in \overline{T} \times \mathcal{R}$.

For any $(t,a) \in \mathcal{T} \times \mathcal{R}$ we may similarly to the inductive case consider the posterior distribution $\rho_{(t,a)}$ defined by

$$\frac{d\rho_{(t,a)}}{d\pi}(t', a') = \frac{\mathbb{1}(t' \in \Delta_t) \mathbb{1}(a' = a)}{\pi\bigl(\Delta_t \times \{a\}\bigr)},$$

but we may also consider $\delta_{(m(\overline{\Delta}_t), a)}$, which is such that $r_i\{[m(\overline{\Delta}_t), a]\} = r_i[(t,a)]$, $i = 1, 2$, whereas only $\rho_{(t,a)}(r_1) = r_1[(t,a)]$, while

$$\rho_{(t,a)}(r_2) = \frac{1}{|\overline{T} \cap \Delta_t|} \sum_{t' \in \overline{T} \cap \Delta_t} r_2[(t', a)].$$

We get

$$\begin{aligned}\mathcal{K}(\rho_{(t,a)},\pi) &= -\log\bigl[\overline{L}(\Delta_t)\bigr] + 2^h\log\bigl(|\mathcal{Y}|\bigr)\\ &\leq \log\bigl(|\overline{T}|\bigr) + 2^h\log(|\mathcal{Y}|) = \mathcal{K}(\delta_{[m(\overline{\Delta}_t),a]},\pi)\\ &\leq h\log\bigl[(k+1)N+1\bigr] + 2^h\log(|\mathcal{Y}|),\end{aligned}$$

whereas we had no such uniform bound in the inductive case. Similarly to the inductive case

$$\pi\bigl[\exp(-\lambda r_1)\bigr] = \sum_{t\in T}\overline{L}(\Delta_t)\prod_{c\in\{0,1\}^h}\Biggl[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Bigl(-\lambda\bigl[b^t(c)-b^t_y(c)\bigr]\Bigr)\Biggr].$$

Moreover, for any $\theta\in\Theta$,

$$\begin{aligned}\pi\bigl\{\exp\bigl[-\lambda r_1 + \xi\rho_\theta(m')\bigr]\bigr\} &= \pi\bigl\{\exp\bigl[-\lambda r_1 + \xi m'(\cdot,\theta)\bigr]\bigr\}\\ &= \sum_{t\in T}\overline{L}(\Delta_t)\prod_{c\in\{0,1\}^h}\Biggl[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Bigl(-\lambda\bigl[b^t(c)-b^t_y(c)\bigr]+\xi\overline{b}(\theta,c)\Bigr)\Biggr].\end{aligned}$$

The bound for the transductive counter part to Theorem 1.39 (page 63), obtained as explained page 102, can be computed as in the inductive case, from these two partitions functions and the above entropy estimates.

Let us mention eventually that, using the same notations as in the inductive case,

$$\begin{aligned}&\pi_{\exp(-\lambda r_1)}\bigl[f_\cdot(X_{N+1})=y'\bigr]\\ &\quad= \Biggl\{\sum_{t\in T}\prod_{c\in\{0,1\}^h}\Biggl[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Bigl(-\lambda\bigl[b^t(c)-b^t_y(c)\bigr]\Bigr)\Biggr]\\ &\qquad\times\sum_{c\in\{0,1\}^h}\overline{L}\bigl[\Delta_t(c)\bigr]\frac{\sum_{y\in\mathcal{Y}}\mathbb{1}(y=y')\exp\bigl\{-\lambda\bigl[b^t(c)-b^t_y(c)\bigr]\bigr\}}{\sum_{y\in\mathcal{Y}}\exp\bigl\{-\lambda\bigl[b^t(x)-b^t_y(c)\bigr]\bigr\}}\Biggr\}\\ &\qquad\qquad\times\Biggl\{\sum_{t\in T}\overline{L}(\Delta_t)\prod_{c\in\{0,1\}^h}\Biggl[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Bigl(-\lambda\bigl[b^t(c)-b^t_y(c)\bigr]\Bigr)\Biggr]\Biggr\}^{-1}.\end{aligned}$$

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale sensitive dimensions, uniform convergence and learnability, J. of ACM **44**(4):615-631, 1997.

[2] J. Y. Audibert, Aggregated estimators and empirical complexity for least square regression, *Ann. Inst. H. Poincaré Probab. Statist.*, **40**, 6, pp. 685-736, 2004.

[3] J. Y. Audibert, *PAC-Bayesian Statistical Learning Theory*, PhD, University Paris 6, June 2004.

[4] A. Barron (1987) Are Bayes Rules Consistent in Information ? *Open Problems in Communication and Computation, T. M. Cover and B. Gopinath Ed.*, Springer Verlag 1987.

[5] A. Barron and Y. Yang (1995) Information Theoretic Determination of Minimax Rates of Convergence, *Ann. Statist.* **27** (1999), no. 5, 1564–1599.

[6] A. Barron, L. Birgé and P. Massart, (1995) Risk bounds for model selection via penalization, *Probab. Theory Related Fields*, **113** (1999), no. 3, 301–413.

[7] G. Blanchard, The "progressive mixture" estimator for regression trees, *Annales de l'I.H.P.*, **35**(6):793-820, 1999.

[8] G. Blanchard, A new algorithm for Bayesian MCMC CART sampling, *preprint*, 2000.

[9] G. Blanchard, *Mixture and aggregation of estimators for pattern recognition. Application to decision trees [Méthodes de mélange et d'agrégation d'estimateurs en reconnaissance de formes. Application aux arbres de décision.]* in English with an introduction in French, PhD dissertation, Université Paris XIII, January 2001.

[10] L. Birgé and P. Massart (1997) From model selection to adaptive estimation, *Festschrift for Lucien Le Cam*, 55–87, Springer, New York.

[11] L. Birgé and P. Massart (1995) Minimum contrast estimators on sieves, Bernoulli **4** (1998), no. 3, 329–375.

[12] L. Birgé and P. Massart, A generalized C_p criterion for Gaussian model selection, *preprint*, 2001.

[13] L. Birgé and P. Massart, Gaussian model selection, *J. Eur. Math. Soc.*, 2001.
`http://www.springer.de/link`

[14] A. Blum and J. Langford, PAC-MDL Bounds, *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Com-*

putational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings, Lecture Notes in Computer Science 2777, Springer 2003, pp. 344-357.

[15] O. Catoni, Laplace transform estimates and deviation inequalities *Ann. Inst. H. Poincaré Probab. Statist.* **39** (2003), no. 1, 1–26.

[16] O. Catoni, Data compression and adaptive histograms. In F. Cucker and J.M. Rojas, editors, *Foundations of Computational Mathematics, Proceedings of the Smalefest 2000*, pages 35–60. World Scientific, 2002.

[17] O. Catoni, Statistical learning theory and stochastic optimization, *Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*, Springer Lecture Notes in Mathematics, **1851**, pp 1 - 270, 2004.

[18] N. Cristianini and J. Shawe Taylor, *An introduction to Support Vector Machines and other kernel based learning methods,* Cambridge University Press, 2000.

[19] M. Feder and N. Merhav, Hierarchical Universal Coding, *IEEE Trans. Inform. Theory*, vol 42, no 5, Sept, 1996.

[20] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning*, Springer, New York, 2001.

[21] J. Langford and D. McAllester, Computable Shell Decomposition Bounds, *Journal of Machine Learning Research*, **5**, 2004, pp. 529-547 (communicated at COLT 2000).

[22] J. Langford and M. Seeger, Bounds for Averaging Classifiers, *technical report CMU-CS-01-102*, Carnegie Mellon University, jan. 2001, `www.cs.cmu.edu/~jcl`.

[23] J. Langford, M. Seeger and N. Megiddo, An Improved Predictive Accuracy Bound for Averaging Classifiers, *International Conference on Machine Learning* **18** (2001), 290-297.

[24] N. Littlestone and M. Warmuth, Relating data compression and learnability. *Technical report,* University of California, Santa Cruz, 1986.

[25] D. A. McAllester, Some PAC-Bayesian Theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, 230–234 (electronic), ACM, New York, 1998;

[26] D. A. McAllester, PAC-Bayesian Model Averaging, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, 164–170 (electronic), ACM, New York, 1999;

[27] C. McDiarmid, Concentration *Probabilistic Methods for Algorithmic Discrete Mathematics*, Habib M., McDiarmid C. and Reed B. Eds.,

Springer, 1998.

[28] E. Mammen and A. Tsybakov, Smooth discrimination analysis. *Ann. Statist.* **27** (1999), 1808-1829.

[29] B. Y. Ryabko, Twice-universal coding, *Probl. Inform. Transm.*, vol 20, no 3, pp. 24-28, July-Sept 1984.

[30] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, *Journal of Machine Learning Research* **3** (2002), 233–269.

[31] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, *Informatics Research Report, Division of Informatics, University of Edinburgh* EDI-INF-RR-0094 (2002), 1–42. `http://www.informatics.ed.ac.uk`

[32] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, *IEEE Trans. Inform. Theory* **44** (1998), no. 5, 1926–1940.

[33] J. Shawe-Taylor and N. Cristianini, On the generalization of soft margin algorithms, *IEEE Trans. Inform. Theory* **48** (2002), no. 10, 2721–2735.

[34] A. Tsybakov, Optimal aggregation of classifiers in statistical learning, *Ann. Statist.* **32**(2004), no. 1, 135-166.

[35] A. Tsybakov, S. Van de Geer, Square root penalty: adaptation to the margin in classification and in edge estimation, *Ann. Statist.*, **33**, pp 1203-1224.

[36] S. Van de Geer, *Applications of empirical process theory.* Cambridge Series in Statistical and Probabilistic Mathematics, 6. Cambridge University Press, Cambridge, 2000. xii+286 pp.

[37] V. N. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.

[38] J.-P. Vert, Double mixture and universal inference, *preprint*, 2000, `http://cg.ensmp.fr/~vert/publi/` .

[39] J.-P. Vert, Adaptive context trees and text clustering, *IEEE Trans. Inform.Theory*, **47** (2001), no. 5, 1884-1901.

[40] J.-P. Vert, Text categorization using adaptive context trees, *Proceedings of the CICLing-2001 conference*, A. Gelbukh (Ed.), LNCS 2004, Springer-Verlag, pp. 423-436, 2001.

[41] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, The Context-Tree Weighting Method: Basic Properties, *IEEE Trans. Inform. Theory*, vol 41, no 3, May, 1995.

[42] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, Context Weighting for General Finite-Context Sources, *IEEE Trans. Inform. Theory*, vol 42, no 5, Sept, 1996.

[43] T. Zhang, From epsilon-entropy to KL-entropy: analysis of minimum information complexity density estimation, *Ann. Statist.* (2006), to appear.

[44] T. Zhang, Information Theoretical Upper and Lower Bounds for Statistical Estimation, *IEEE. Trans. Inform. Theory*, (2006), to appear.